Journal of Educational Psychology

VOLUME 66 NUMBER 1

FEBRUARY 1974

AMERICAN PSYCHOLOGICAL ASSOCIATION

EDITOR

JOANNA WILLIAMS, *Teachers College, Columbia University*

ASSOCIATE EDITOR

SAMUEL BALL, *Educational Testing Service*

ADVISORY EDITORS

PAUL R. AMMON, *University of California, Berkeley*
JERE BROPHY, *University of Texas, Austin*
C. VICTOR BUNDERSON, *Brigham Young University*
LEONARD S. CAHEN, *Educational Testing Service*
LOIS-ELLIN DATTA, *National Institute of Education*
FREDERICK B. DAVIS, *University of Pennsylvania*
FRANCIS J. DIVESTA, *Pennsylvania State University*
ROBERT EBEL, *Michigan State University*
DAVID ELKIND, *University of Rochester*
LAWRENCE T. FRASE, *Bell Laboratories*
N. L. GAGE, *Stanford University*
ROBERT GAGNÉ, *Florida State University*
ROBERT GLASER, *University of Pittsburgh*
MIRIAM L. GOLDBERG, *Teachers College, Columbia University*
EDMUND W. GORDON, *Teachers College, Columbia University*
VERNON C. HALL, *Syracuse University*
ARTHUR R. JENSEN, *University of California, Berkeley*
PHYLLIS A. KATZ, *City University of New York*
HERBERT J. KLAUSMEIER, *University of Wisconsin—Madison*
ROBERT E. LANA, *Temple University*
GERALD S. LESSER, *Harvard University*
JOEL R. LEVIN, *University of Wisconsin—Madison*
FREDERICK J. MCDONALD, *Educational Testing Service*
WILBERT J. MCKEACHIE, *University of Michigan*
FRANK B. MURRAY, *University of Delaware*
LAUREN B. RESNICK, *University of Pittsburgh*
WILLIAM D. ROHWER, JR., *University of California, Berkeley*
ROBERT ROSENTHAL, *Harvard University*
ERNST Z. ROTHKOPF, *Bell Laboratories*
S. JAY SAMUELS, *University of Minnesota*
SEYMOUR B. SARASON, *Yale University*
MELVYN I. SEMMEL, *Indiana University*
IRVING SIGEL, *Educational Testing Service*
HARRY SILBERMAN, *University of California, Los Angeles*
RICHARD E. SNOW, *Stanford University*
GEORGE G. THOMPSON, *Ohio State University*

JOURNAL STAFF

HAROLD P. VAN COTT, *Managing Editor*
ANITA DEVIVO, *Executive Editor and General Manager*
DEBORAH S. FELDER, *Supervising Technical Editor*
ANN G. CORBETT, *Technical Editor*
FORREST T. MULLINS, *Manager, Publications Business Service*
ANNE REDMAN, *Subscription Manager*

Published bimonthly in one volume per year, six issues (February, April, June, August, October, and December), by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D. C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D. C., and additional mailing offices.

The **Journal of Educational Psychology** publishes original investigations and theoretical papers dealing with problems of learning and teaching and with the psychological development, relationships, and adjustment of the individual. Preference is given to studies of the more complex types of behavior, especially in or relating to educational settings. Journal articles pertain to all levels of education and to all age groups.

Manuscripts: Submit manuscripts in duplicate to the Editor, Joanna Williams, P.O. Box 51, Teachers College, Columbia University, New York, New York 10027, according to instructions provided elsewhere in the Journal.

Copyright and Permissions: All rights reserved. Written permission must be obtained from the American Psychological Association for copying or reprinting tables, figures, or text of more than 500 words. Permissions are normally granted contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of $10 per page of APA material. Permission fees are waived for authors who wish to reproduce their own articles, or when material is reproduced in limited quantities for instructional purposes. Address requests for permission to the Permissions Office, American Psychological Association, 1200 Seventeenth Street, N. W., Washington, D.C. 20036.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1974: $24 domestic, $25 foreign, $5 single issue. APA member rate: $10. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their article from the printer when they receive galley proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Department, APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., 175 Fifth Avenue, New York, New York 10010; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilm, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. Undelivered copies resulting from address changes will not be replaced; second-class mail is not forwarded unless subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

EDUCATIONAL AND PSY... GOVT OF W. BENGAL LIBRARY CALCUTTA-16

Volume 66 February 1974 Number 1

Selective Facilitation and Interference in Retention of Prose 1
GORDON H. BOWER

Coeducation, Values, and Satisfaction with School 9
N. T. FEATHER

Remediation of Learning Problems among the Disadvantaged 16
SEYMOUR FESHBACH AND HOWARD ADELMAN

Attention and Reading Achievement in First-Grade Boys and Girls 29
S. JAY SAMUELS AND JAMES E. TURNURE

Relationships of State and Trait Anxiety to Concept-Learning Performance .. 33
JOEL MEYERS AND ROY MARTIN

Effects of Position and Type of Question on Learning from Prose Material: Interaction of Treatments with Individual Differences 40
RICHARD J. SHAVELSON, DAVID C. BERLINER, MICHAEL M. RAVITCH, AND DAVID LOEDING

Note on Crano, Kenny, and Campbell's "Does Intelligence Cause Achievement?" 49
JEAN L. DYER AND LOUISE B. MILLER

Initial Level of Student Evaluation of Instruction as a Source of Influence on Instructor Change after Feedback 52
HAGOP S. PAMBOOKIAN

Sequence Effects and Reading Time in Programmed Learning 57
JAMES W. DYER AND RAYMOND W. KULHAVY

Effects of Test Score Feedback on Immediately Subsequent Test Performance .. 62
BRENT BRIDGEMAN

Effects of Models of Creative Performance on Ability to Formulate Hypotheses .. 67
NORMAN FREDERIKSEN AND FRANKLIN R. EVANS

Elaborative Prompt Effects in Children's Paired-Associate Learning: Developmental and Population Comparisons 83
JANET D. FREDERIKSEN AND WILLIAM D. ROHWER, JR.

Monetary Incentive Effects upon Incidental Learning during an Instructional Task 90
STEPHEN WOLK AND JOSEPH DUCETTE

Immediate and Delayed Retention Effects of Interspersing Questions in Written Instructional Passages 96
JOHN R. BOKER

(Continued on next page)

809
16/9/74
Bureau of Educl. & Psyl. Res...
(N. C. E. R. T.)
Date 886

Interaction of Level I and Level II Abilities with Race and Socioeconomic Status 99
ARTHUR R. JENSEN

Stimulus and Response Variables in Children's Learning of Grapheme–Phoneme Correspondences 112
GEORGE MARSH, PETER DESBERG, AND LAWRENCE KENT FARWELL

Biasing Effects of Diagnostic Labels and Sex of Pupil on Teachers' Views of Pupils' Mental Health 117
PHYLLIS F. HERSON

Personality Traits Associated with Effective Teaching in Rural and Urban Secondary Schools 123
KENNETH D. MATTSSON

Sex Differences in the Organizational Assimilation of Beginning Graduate Students in Psychology 129
JOHN E. NEWMAN

Formal Discipline Revisited: Affective Assessment and Nonspecific Transfer .. 139
JOSEPH F. RYCHLAK, NGUYEN DUC TUAN, AND WILLIAM E. SCHNEIDER

Construct Validity of Test Items Measuring Acquisition of Information from Line Graphs 152
JAY R. PRICE, VICTOR R. MARTUZA, AND JAMES H. CROUSE

List of Manuscripts Accepted 122

Announcement 151

SUGGESTIONS TO CONTRIBUTORS

In preparing articles for publication in the Journal, authors are strongly advised to follow the general directions given in the *Publication Manual of the American Psychological Association* (1967 Revision). All articles must be preceded by an abstract of 100–120 words. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833). Special attention should be given to the section on the preparation of the references (p. 50). In reference lists, give journal titles in full; do not abbreviate. *All copy must be double spaced, including the references.* All manuscripts should be submitted in *duplicate*, one of which should be an original typed copy. Since the reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the identity of the author or his affiliation should be on separate pages. Every effort should be made by the author to see that the manuscript itself contains no clues as to his identity. Dittoed and mimeographed copies are not acceptable and will not be considered. Original figures must be fully prepared for reproduction according to the instructions in the *Manual*. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail and to check carefully the typing of the final copy.

Journal of Educational Psychology
1974, Vol. 66, No. 1, 1-8

SELECTIVE FACILITATION AND INTERFERENCE IN RETENTION OF PROSE[1]

GORDON H. BOWER[2]

Stanford University

A prose passage may be described at the level of asserted relations among general semantic categories or at the level of predications about specific details (e.g., names, dates, places). This experiment used interpolated learning to facilitate retention of the *conceptual macrostructure* of an originally learned passage while simultaneously interfering with retention of the originally learned passage's *detailed microstructure*. The subjects originally learned a short biography; the experimental subjects then learned two more biographies of similar conceptual format which had one third of the details changed. Later, cued recall of changed details of the originally learned passage showed retroactive interference and many intrusion errors. Free recall of the originally learned passage showed depressed recall of changed details and enhanced recall of unchanged details, but equal facilitation in recall of the conceptual macrostructure.

This experiment investigates interference processes in retention of meaningful prose. Although earlier investigations of retroactive interference with prose materials typically found weak effects, strong effects have been obtained recently in experiments that have taken more care in arranging optimal interfering conditions and in analyzing appropriate aspects of recall in accordance with interference principles established in the typical "nonprose" experiments (e.g., Anderson & Myrow, 1971; Crouse, 1971; Myrow & Anderson, 1972). The present experiment extends this analytic approach to further interference phenomena in prose learning.

Previous investigations suggest that one can devise interpolated texts which when learned can produce either retroactive interference, retroactive facilitation, or "no effect" with respect to the retention of an originally learned passage. To illustrate, suppose that an originally learned passage is a short biography of a fictitious poet named John Payton and it contains the complex sentence, "Payton's father worked as a blacksmith but he died of diphtheria when John was only five years old." Consider a counterpart sentence that occurs in the interpolated passage, another biography about another fictitious poet, Robert Fowler: "Fowler's father worked as a blacksmith but he died of lung cancer when Robert was only three years old." Comparing these sentences, some details have remained the same whereas others have changed. Assuming initial learning of the first sentence, the second sentence should help the subject retain the original fact that the poet's father was a blacksmith, because that predicate construction is repeated for the main character. However, the disease that killed the father and the son's age when his father died are details that have been changed. For these, one expects response competition and retroactive interference when test questions explicitly probe for recall of their originally learned values. These two cases illustrate retroactive facilitation and retro-

[1] This research was supported by grant MH13950-06 to the author from the National Institute of Mental Health. Robert Rothbart and John Anderson assisted in the collection and analysis of these data. The author wishes to thank James Crouse for making his experimental materials available to him and Richard Anderson for several unpublished (preprinted) manuscripts.

[2] Requests for reprints should be sent to Gordon H. Bower, Department of Psychology, Stanford University, Stanford, California 94305.

active interference, respectively. The neutral or noninterference condition would have, in the interpolated passage, a sentence that mentioned neither the subjects nor the predicates of the original passage. A given interpolated passage will have certain proportions of facilitating, ineffective, and interfering constructions with respect to an originally learned text, and its net effect upon originally learned retention should vary systematically with the proportions of the three constructions among the test items for originally learned retention. Our investigation attempts to validate this commonsense analysis.

A second issue we investigated is the recall of the *conceptual macrostructure* of the passage; this refers to the main conceptual categories and relations (or gist) of the passage without regard to the detail carried by names, dates, places, and numbers. Thus, the macrostructure of the two sentences above is that "the main character's father had *some* occupation but he died of *some* disease when the character was only *some* years old." The "somes" denote variables into which different instantiating details are inserted to produce a facilitating or interfering sentence for the interpolated passage. This formulation suggests that recall of the categories of an originally learned passage may be enhanced by an interpolated passage that alters and interferes with memory of many of the originally learned details. The following experiment attempts to create this pattern, that is, retroactive inhibition in recall of the detailed microstructure of the passage, while at the same time enhancing retention of its conceptual macrostructure.

We do this by use of unprompted or "free" recall, examining the conceptual relations recalled as well as the accuracy of the details instantiating those categories. Provided that the subject guesses when uncertain among competing answers, we should find retroactive facilitation in his recall of the conceptual macrostructure of the passage. In addition, facilitation or interference should be observed for recall of specific details of the originally learned passage, depending on whether these are repeated or changed in the interpolated text. These "detail-specific" effects should be as apparent in accuracy of cued-recall as in free-recall measures of originally learned retention. The following experiment tested these interlocking expectations.

Method

Two groups of undergraduates (13 in each group, approximately half male and half female) served as subjects and were paid $1.75 for one hour's service. Each subject learned three successive text passages and then received a final retention test over the initial two passages. All subjects originally learned the biography of a fictitious poet, John Payton. This passage, adapted from Crouse (1971), was arranged into 15 distinct sentences totaling 173 words. The first 4 sentences are illustrative:

> John Payton was one of the finest poets England has ever known. Payton was born in Northshire at the end of October, 1810. His father was a servant who worked in the nearby town of Blackrock. When Payton was only five years old, his father was killed by a robber.

These passages were shown to small groups of from 2 to 4 subjects by means of an overhead projector. The individual sentences were shown one every 5 seconds by sliding a cardboard "mask" down a translucent sheet exposing the sentences one by one. The subjects were told to memorize each sentence but also to relate it to the overall text. After one study trial and a 10-second pause, the same originally learned passage was studied again. Following the second study trial, subjects read aloud in unison a slide of random digits for 20 seconds (to minimize short-term memory), and then they were asked to free recall the 15 sentences of the passage studied. Recall instructions were somewhat vague: subjects were told that we were interested in how much they could recall verbatim (hence the 15 numbers down one side of the recall sheet for the corresponding sentences); but they were further urged to reproduce the gist or substance of any facts they could not recall verbatim and also to write any fact they could recall regardless of whether they knew its order of presentation. The subjects had 6 minutes to write their free recall, and all finished by that time. We will refer to this as the immediate free-recall test of original learning.

After collecting the free-recall sheets, the experimenter distributed the cued-recall sheet for the originally learned passage. This asked 20 questions (having one–three word answers) regarding specific details of the passage just studied. For example, for the Payton story, 1 question asked for the occupation of Payton's father, another asked for Payton's year of birth, another asked how old Payton was at the time his father died, etc. The subjects completed these cued-

recall sheets in 2.5 minutes. We will refer to this as the cued-recall test for the originally learned passage.

Upon arrival at the lab, subgroups of subjects had been assigned randomly to the experimental or control condition (subgroups of from two to four subjects in these two conditions were run in random alternation). The two treatment groups received different interpolated passages, each learning and recalling two separate interpolated passages (to enhance the hoped-for retroactive interference). The experimental subjects learned two more biographies which were very similar to the one for John Payton except that the main character had a different name (Robert Fowler, then Richard Hughes) and 22 specific details of the originally learned passage were changed. To illustrate the changes, the first four sentences of the "Robert Fowler" passage were identical to those for John Payton except for the details that Fowler was born in Hampstead in 1795 of a blacksmith who was killed by a wolf when Fowler was eight years old. The third, the "Richard Hughes" passage, changed the same factual details as did the second passage. So in comparing these passages to the originally learned passage, one can pinpoint which details have remained the same and which have changed. Control subjects learned two unrelated passages of similar length and difficulty, the first describing the collections of the fictitious King Library, the second describing the geography and inhabitants of the fictitious island of Karisoon. These passages were also adapted from Crouse (1971).

Learning of each interpolated passage followed the same procedure as the originally learned passage: two presentations of the single sentences of the passage, followed by 20-second digit reading, followed by a 6-minute written free recall, and then a 2.5-minute cued-recall test with 20 questions regarding details. For experimental subjects, the 20 questions on these interpolated tests were identical to those used for the originally learned passage except that the name of the poet was changed.

Following the cued-recall test on the last interpolated passage, retention of the originally learned passage was tested, yielding the data of primary interest. First, free recall was requested of "the first passage, the one about John Payton." After 6 minutes, the subjects then received the cued-recall test for the originally learned passage (for 2.5 minutes), asking the same questions again. Because of time limitations, we were unable to test retention in both ways for the interpolated passages. However, a final cued recall was obtained for the first interpolated passage, List 2.

Protocol Scoring

Each subject contributed four free-recall protocols and five cued-recall protocols. Cued recall was scored as correct or wrong and, if wrong, whether or not an intrusion error from an incorrect list occurred. A lenient criterion for gist recall was used. Free-recall scoring introduced several problems. We had to decide what were the parts of conceptualizations expressed in the text. These were basically of two types: (*a*) attributive or modificatory statements (e.g., "disease was cancer," "John's age was five years") or (*b*) nouns serving as case-relation arguments in relation to the main verb—either as agent, recipient, time or location of the action (verb). For instance, the sentence, "his father was killed by a robber," has the two argument–relation pairs stipulating "robber–killed" (agent–act) and "killed–father" (act–recipient). Although one might quibble over our exact segmentation of the sentences into "idea units," it is doubtful whether the overall pattern of the results would be much affected by these identifications (see Paul, 1959).

By this means, we identified 63 "idea units" in the originally learned passage; of these, 22 had been changed in detail in both of the interpolated passages (learned by the experimental subjects), whereas 41 had remained the same throughout the originally learned passage and both interpolated lists. Recalls of these two kinds of originally learned units were tabulated separately. Each idea unit of the originally learned list was scored according to whether it was recalled as a specific detail (e.g., that Payton's age was five years at the time his father was killed) and according to whether the right general kind of fact was recalled (e.g., mentioning Payton's age at the time his father was killed). We called this latter the subject's general fact recall and the former his specific fact recall. These are correlated measures, since correct recall of a specific fact entails recall of the corresponding general fact. However, the scoring scheme permits a general fact to be recalled without its specific detail being correct (due to intrusions or simple mistakes). Scoring was lenient with respect to paraphrases and synonyms but nonetheless was quite reliable for two judges (scoring was not "blind" for either judge).

Results

Original Learning

The degree of original learning for the three successive lists is shown by the immediate cued-recall proportions in Table 1. The experimentals and controls did not differ reliably in immediate recall on any list. This was true for free recall as well as for cued recall. There was a general practice effect from List 1. Although the experimentals are slightly inferior to the controls on Lists 2 and 3, a result possibly reflecting negative transfer, that inference is unwarranted due to a covarying (small) difference on List 1 performance and also to

TABLE 1

CUED-RECALL PROPORTIONS ON THE IMMEDIATE TEST FOR LISTS 1, 2, AND 3 AND CUED-RECALL PROPORTIONS AND INTRUSION PROPORTIONS ON THE DELAYED TEST FOR LISTS 1 AND 2

Condition/test	List 1		List 2		List 3	
	Experimental	Control	Experimental	Control	Experimental	Control
Cued recall						
Immediate	.54	.59	.71	.78	.68	.75
Delayed	.40	.59	.59	.75		
Intrusion						
Delayed	.10	—	.12	—		

the fact that the Lists 2 and 3 which were learned by the two groups differed, so there is no assurance that the lists were of comparable difficulty.

Forgetting: Cued Recall

The cued-recall loss scores for Lists 1 and 2 are calculated by subtracting the delayed from the immediate cued-recall proportions in Table 1 (time did not permit a delayed test on List 3). For List 1, the experimental subjects had a specific loss of about 13.9%, whereas the controls forgot .4%. An overall analysis of variance was carried out on transformed (arcsin) recall proportions with experimental–control interpolation as a between-subjects factor and immediate–delayed testing as a within-subjects factor (see Winer, 1962, p. 302 ff.). As expected, that analysis revealed significant forgetting ($F = 14.6$, $df = 1/24$, $p < .01$), with reliably more forgetting by the experimental than the control subjects (interaction, $F = 11.9$, $df = 1/24$, $p < .01$). Similar conclusions stem from comparing forgetting of List 2 items by the two groups ($p < .01$).

A revealing statistic in Table 1 is the proportion of intrusions of competing responses from interfering lists learned by the experimental subjects (e.g., recalling a birthdate from List 2 or 3 while trying to recall the one stated in List 1). These competitive intrusions were a major determinant of the differential loss in the originally learned passage for the experimental subjects.

Free Recall

Specific facts. Free recall of specific factual details was examined first. The proportions of specific originally learned facts recalled on the immediate and delayed tests are shown in Table 2, divided according to changed versus unchanged facts and according to experimental versus control treatments. For the controls, the changed versus unchanged set refers to those items in the originally learned passage that were changed or not changed in the interpolated passages learned by the experimental subjects. The total proportions for the changed and unchanged scores combined are shown for each cell, and the changes from the immediate- to the delayed-recall scores are easily computed.

These data allow several interesting comparisons whose discussion can begin with the total scores. Much as with earlier studies, these totals revealed no net retroactive interference. An overall analysis of variance on total free-recall proportions (arcsin total) yielded no effect due to treatment groups or to the interaction of treatment with retention interval ($F = 1.44$, $df = 1/24$, $p > .10$). Surprisingly, there was even a slight *increase* in the total facts recalled between the immediate and the delayed retention test.

However, the detailed story of what happened is told by separately considering recall of the changed versus the unchanged specific facts. Relative to the controls, the experimental subjects forgot changed details while increasing their recall of unchanged (repeated) details. An analysis of variance was performed on the transformed proportions (see Table 2) with experimental versus control as a between-subjects factor, and changed versus unchanged facts and immediate versus delayed tests as two

within-subjects factors (see Winer, 1962, p. 319 ff.). The effects of chief interest appear as interactions. First, the changed–unchanged factor interacts with immediate versus delayed testing ($F = 104$, $df = 1/24$, $p < .01$), reflecting an increase in the unchanged versus changed difference over the retention interval. Second, there was a significant triple interaction of the above two factors with the experimental versus control contrast ($F = 34$, $df = 1/24$, $p < .01$). As expected, then, the results show retroactive interference in recall of changed details and facilitation for unchanged (repeated) details, with the changes being greater for the experimental than for the control subjects. These specific contrasts were confirmed by significant t tests. One main effect was significant in the overall analysis—unchanged facts were recalled better than changed facts, even on the immediate test—but that is a *materials construction* artifact of no particular interest. No other effects were significant in the overall analysis on free recall of specific facts. Part of the loss in recall of changed details was due to intrusions of interpolated details while the subject was attempting recall of the originally learned detail (average of 1.77 per subject).

General facts. A *general fact* is said to be recalled if the subject recalls the correct general category of a thing or relation between things, though possibly with the wrong detail. Free-recall proportions for such general facts are shown in Table 3 for immediate and delayed tests, the general facts divided according to whether their details were changed or unchanged between the originally learned and the interpolated passages.

TABLE 2

FREE-RECALL PROPORTIONS FOR SPECIFIC FACTS ON IMMEDIATE AND DELAYED TESTS

Test/specific facts	Experimental	Control
Immediate		
Changed	.48	.53
Unchanged	.61	.68
Total	.55	.63
Delayed		
Changed	.34	.51
Unchanged	.76	.73
Total	.62	.66

TABLE 3

FREE-RECALL PROPORTIONS FOR GENERAL FACTS ON IMMEDIATE AND DELAYED TESTS

Test/general facts	Experimental	Control
Immediate		
Changed	.62	.67
Unchanged	.66	.72
Total	.64	.70
Delayed		
Changed	.76	.71
Unchanged	.80	.77
Total	.79	.75

These general-fact-recall proportions were used in an analysis of variance similar to that described above, with one between-subjects factor and two within-subjects factors. The chief effects of interest are the overall increase in recall scores from the immediate to the delayed test (for retention interval, $F = 21.68$, $df = 1/24$, $p < .01$) and the fact that this increase was larger for experimental than for control subjects (interaction, $F = 5.40$, $df = 1/24$, $p < .05$). The changed versus unchanged factor did not enter into any of these significant interactions, although it was associated with a significant main effect: unchanged items were recalled slightly better than changed items even initially ($F = 6.36$, $df = 1/24$, $p < .05$); this is the same materials construction artifact noted before. The difference scores (delayed minus immediate free-recall proportions) make transparent the pattern of significant outcomes: experimental subjects improved their free recall of general facts more than control subjects, and this improvement was as large for general facts with changed details as for those having unchanged details. We may therefore conclude that improvement in recall of the conceptual macrostructure is approximately the same whether the underlying details are changed or kept the same from the original to the interpolated texts.

DISCUSSION

The results accorded with commonsense expectations: retroactive interference was observed for changed details and retroac-

tive facilitation for unchanged details. Interpolated passages that were similar to the originally learned passage but which had altered details were nonetheless beneficial in facilitating retention of the conceptual macrostructure of the original passage. The interpolated passage boosted free recall of the appropriate *kinds* of facts, while lowering the accuracy of the details recalled. The effect here is similar to that encountered in free recall of categorized word lists: a taxonomic category that is repeated on an interpolated list but with different instances will be thereby enhanced in category recall of the original list but at the expense of a lower score on a conditionalized measure of "items per recalled category" (see Shuell, 1968).

Some discussion is needed regarding limitations on the generality of our result. First, the learning and retention intervals here were short, spanning 35 minutes at most, and similar experiments are required at longer intervals and with larger amounts of learning material. Second, a serious limitation is that the experiment used only one specific biography as the critical originally learned material. One should like to generalize the results to a larger population of materials—and not only to biographies but to other "fact-listing" passages and possibly even to less "fact-oriented" materials. Therefore, replication of the experiment with several different sets of materials would clearly enhance the generality of its principal conclusions.

An issue requiring discussion is whether our procedures oriented the subject far too much to word-for-word or verbatim memorization, stripping the text of its meaningfulness and only thereby creating conditions favorable for demonstrating interference effects in retention. We have serious doubts about the dichotomies implicit in such a remark. Nothing in interference theory restricts its applicability to meaningless material or to "factual details" such as names, dates, locations, and numbers. Simple sentences may be conceived of as being learned by associating concepts corresponding to the subject and predicate of the sentence (see Anderson & Bower, 1973). For example, using such simple sentences as "the uncle shouted an obscene remark," Anderson (1971) showed that following study of such sentences, later recall of the subject term (*uncle*) was cued almost as well by a *paraphrase* of the predicate ("——yelled some dirty words") as by the *verbatim* predicate that had been studied. Moreover, this close correlation in cueing effectiveness of the verbatim predicate and its paraphrased predicate was also found for subjects instructed to learn the original sentences by "rote, verbatim repetition." The implication is that even with instructions designed to promote "nonnormal," rote processing of sentences, the subject nonetheless and willy-nilly established associations between semantic concepts. In another relevant study, Anderson and Carter (1972) found that recall of a subject–predicate construction showed retroactive interference from interpolated learning of a paraphrased predicate paired with a different subject (the originally learned predicate was the recall cue for the originally learned retention test). Such interference must have been at a semantic level. Such evidence indicates that interference principles apply to retention of meaningful text passages.

A fourth issue requiring discussion concerns the clarity and utility of the distinction between the conceptual macrostructure of a passage and its detailed microstructure. This distinction currently rests on linguistic intuitions. Even formal systems for analyzing the semantical structure of text such as those of Crothers (1972) and Fredericksen (1972) only try to describe explicitly our intuitions regarding successive "levels of abstraction" from a specific text. Starting from a detailed statement (e.g., "Mary washed her pet cat"), we may generalize it by substituting a more general category for each content word or by deleting a qualifier (e.g., "a girl cleaned her pet"); alternatively, we may generate an interfering specific statement by replacing the original instances with others. If these substitutes are within the same category (e.g., "Mary brushed her pet dog"), then the same categorical relationships have been repeated and these connections should be strengthened in memory. If the substituted instances come from different categories

(e.g., "Mary cleaned her bedroom"), then interference or *unlearning* would be expected of the original conceptual (general) relation. Admittedly, assignment of particular statement pairs to different levels of abstraction is not always clear, nor is the decision as to what is the same versus a different conceptual relation. Nonetheless, the decisions are sharp in many comparisons (e.g., names, dates, places in biographies), and it is an important analysis to aim for. Why? Because, following Bartlett (1932), most of us believe that text recall is initially a matter of the person remembering the general gist or an "abstract outline" of a text, and then fleshing out the details of this abstract (partly by "reconstruction"). Our biography-recall data illustrate in a simple, transparent context how repetition of an abstract outline facilitates recall of at least the right general kinds of facts. Presumably, the same facilitation could be shown for materials other than listings of specific names, dates, numbers, and places. Presumably, also, one should be able to *interfere* with recall of the conceptual macrostructure by breaking the originally learned conceptual relations with the interpolated passage. These are matters for future research.

Finally, let us consider a practical matter, namely, interference potentialities in learning school materials. A critic might point out that in the present experiment, retroactive interference was produced only by highly similar passages, and even then the magnitude of the forgetting was not dramatic. Is such retroactive interference, then, a phenomenon likely to be restricted to only a small portion of what a student learns? The critic may ask, "After all, how often does a curriculum teach a student totally different answers to the same question?"

The answer, of course, is "practically never," but the critic's charge plays upon a deceptive indefiniteness in our notions of "the same question" and "different answers." Suppose that memory for a simple proposition depends upon associating the semantic concept underlying the subject (A) to that underlying the predicate (B) so that the associative structure is A–B. By such analyses, then, interference effects should arise whenever knowledge requires a "multiple listing" of predicates that apply to the same concept or similar concepts (see Anderson & Bower, 1973). One example would be a zoological classification in which a species is characterized by a list of salient attribute–value pairs (e.g., heart type, vertebrate–invertebrate, habitat, color, size, etc.) and different species are defined by a conjunction of different values on the same attributes. A second example would be temporal conjunctions defining historical or contemporaneous events. A schematic example would be, "When A was King and B was his General, War C was fought. But when A′ was King and D was the General, War E was fought." Confusion would surely reign when the student sorts out in memory which king went with which general and which pair got involved in which war. A third example where interference may be expected is when the same objects (e.g., different historical characters) are to be rank ordered along two independent dimensions, such as historical date and relative power or senatorial seniority and political influence. To the extent that objects occupy differing locations in the two linear orders, confusions and interference may be expected in learning and keeping separate the two orders.

These examples—and many more of this general type come readily to mind—suggest many opportunities for interference processes to exert a pervasive influence on learning and retention in the usual curriculum. Moreover, these interference effects would seem to apply just as formidably to the retention of "meaningful principles" or rules as it does to "straight facts." First, the distinction between *principles* and *facts* is none too clear anyway. Second, under proper analysis, *rules* would seem to be decomposable into atomic propositions connected by conditionals which, if the indicated analysis has merit, are just as vulnerable to retroactive interference when it is "aimed" at the proper level.

REFERENCES

ANDERSON, J. R., & BOWER, G. H. *Human associative memory*. Washington, D. C.: V. H. Winston, 1973.

ANDERSON, R. C. Encoding processes in the storage and retrieval of sentences. *Journal of Experimental Psychology*, 1971, **91**, 338–340.

ANDERSON, R. C., & CARTER, J. F. Retroactive inhibition of meaningfully-learned sentences. *American Educational Research Journal*, 1972, **9**, 443–448.

ANDERSON, R. C., & MYROW, D. L. Retroactive inhibition of meaningful discourse. *Journal of Educational Psychology Monograph*, 1971, **62**, 81–94.

BARTLETT, F. C. *Remembering*. Cambridge: Cambridge University Press, 1932.

CROTHERS, E. J. Memory structure and the recall of discourse. In J. B. Carroll & R. O. Freedle (Eds.), *Language comprehension and the acquisition of knowledge*. Washington, D. C.: V. H. Winston, 1972.

CROUSE, J. H. Retroactive interference in reading prose materials. *Journal of Educational Psychology*, 1971, **62**, 39–44.

FREDERICKSEN, C. H. Effects of task-induced cognitive operations on comprehension and memory processes. In J. B. Carroll & R. O. Freedle (Eds.), *Language comprehension and the acquisition of knowledge*. Washington, D. C.: V. H. Winston, 1972.

MYROW, D. L., & ANDERSON, R. C. Retroactive inhibition of prose as a function of the type of test. *Journal of Educational Psychology*, 1972, **68**, 303–308.

PAUL, I. H. Studies in remembering: The reproduction of connected and extended verbal material. *Psychological Issues*, 1959, **1** (2), 1–152.

SHUELL, T. J. Retroactive inhibition in free-recall learning of categorized lists. *Journal of Verbal Learning and Verbal Behavior*, 1968, **7**, 797–805.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received February 6, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 9–15

COEDUCATION, VALUES, AND SATISFACTION WITH SCHOOL[1]

N. T. FEATHER[2]

Flinders University of South Australia, Bedford Park, Australia

Boys and girls from the two senior classes in eight Adelaide, Australia, coeducational and single-sex state secondary schools ranked sets of values from the Rokeach Value Survey, first in order of importance for themselves (own values) then in the order they thought their schools would emphasize them (school values). They then completed a modified form of the Cornell Job Description Index and a rating of happiness with school. Factor analysis indicated basic similarities across schools in the ordering of both own and school values, but no factor emerged contrasting average value systems for coeducational versus single-sex schools. Boys in coeducational schools were more satisfied with classmates and teachers than were boys in single-sex schools. Results were related to recent theory and research.

It is commonly argued that coeducation at the secondary school level is necessary to prepare children to take their places naturally in the world of men and women. It is contended that the social environment of the coeducational school would be less artificial than that of the single-sex school, and the adaptations learned in an environment that more accurately mirrored that of the wider social context would better equip children to adjust to the adult world beyond the school. The argument is a compelling one, but surprisingly, very few studies have been conducted to compare the effects of coeducational and single-sex schools on the children attending them.

What sorts of studies have been carried out? In Britain, Dale (1969, 1971) has conducted an extensive program of research into coeducation in which questionnaires were administered to pupils, ex-pupils, and teachers. He found that coeducational schools were generally preferred to single-sex schools by both teachers and students. The school atmosphere was thought to be more congenial in coeducational schools, and students saw their teachers as friendlier and more helpful. Single-sex schools were perceived to involve stricter discipline, and teachers in these schools were seen as more distant. Dale (1969) reported that there was a tendency "for the differences between the attitudes of boys in boys' schools and those of boys in mixed schools, towards their school life, to be less sharp than are the comparable differences between the two groups of girls [p. 232]." He also noted that there was no evidence that the education of the sexes together resulted in a lowering of academic standards. Indeed, as far as boys were concerned, the weight of the evidence seemed to be on the opposite side (Dale, 1962a, 1962b, 1964).

A study recently reported by Jones, Shallcrass, and Dennis (1972) and conducted in New Zealand was less positive toward coeducation. Students in a boys' school, a girls' school, and a coeducational school completed items from a questionnaire used by Coleman (1961). The authors were interested in testing Coleman's suggestion that status in the adolescent society of the coeducational secondary school may depend more upon popularity than upon scholastic or intellectual achievement, with a consequent emphasis upon "rating and dating."

[1] This research was funded by a grant from Flinders University and by the Australian Research Grants Committee.

The author is indebted to Andrew Ellerman for his assistance in data analysis.

[2] Requests for reprints should be sent to N. T. Feather, School of Social Sciences, Flinders University of South Australia, Bedford Park, South Australia 5042.

Hence coeducation may have a stultifying effect on intellectual activities and "may be inimical to *both* academic achievement and social adjustment [Coleman, 1961, p. 51]." Jones et al. did, in fact, find some support for Coleman's suggestion in their analysis of the responses of boys and girls to the questionnaire. Furthermore, differences on most items were larger for the two groups of girls than for the two groups of boys. Thus, as in the studies conducted by Dale, the influence of the different types of school was more evident among girls than among boys.[3]

The present study was designed to provide information about the effects of coeducation in the Australian context, using a limited sample of schools drawn from the Adelaide metropolitan area. It focused upon differences between coeducational and single-sex high schools both in regard to the relative importance assigned by students to different values and in regard to their expressed satisfaction with various aspects of the school situation. The concept of value is a central one in social science (Rokeach, 1968a, 1968b), yet curiously there has been less *empirical* examination of general value differences than there has been of differences in more specific attitudes, possibly because of the lack of suitable instruments for assessing values. There has been no lack of *theoretical* discussion of the value concept, however, and of its relevance to wide areas of social science, including education (e.g., see Rokeach, 1968a, 1971, 1973; Williams, 1971). The present study was one aspect of an extensive program of research on values stimulated in part by Rokeach's analysis (for details see Feather, 1972c).

One might expect coeducation to affect both values and attitudes among students. If Coleman (1961) and Jones et al. (1972) are correct, then students from coeducational schools might regard values relating to social approval and affiliation as more important for self than would students from single-sex schools, and this difference might be especially evident among the girls. Following Dale, one might also expect students from single-sex schools to see their schools as placing more emphasis upon values concerned with discipline and control than would students from coeducational schools. As far as attitudes are concerned, if teachers are friendlier and more helpful in coeducational schools, then one would expect this difference to be evident in students' reported satisfaction with their teachers. Similarly, if the social environment of the coeducational school is seen as richer and more complete, involving, as it does, both sexes, then one would expect coed students to express more satisfaction with their classmates. It was possible to explore these types of questions concerning general values and more specific attitudes by using the Rokeach Value Survey (Rokeach, 1968a, 1968b, 1971, 1973) and a modified form of the Cornell Job Description Index (Smith, Kendall, & Hulin, 1969). These measures were selected because they have had a lot of use in recent years, are well researched, and were appropriate to the subjects involved in the study and to the overall aim of the general survey of which the present study was part.

Method

Subjects and Schools

The present study formed part of an extensive survey involving nearly 3,000 school children attending their last two years of secondary school education (leaving and matriculating classes) in 19 schools in the metropolitan district of Adelaide, Australia (Feather, 1972a, 1972b, 1972c). The present investigation involved 8 of these schools, all of which were administered and funded by the state government. Of these, five were high schools offering a fairly general range of courses. The remaining three high schools concentrated more upon technical and commercial courses. The number of subjects from the various schools were as follows: *coeducational schools*—General High School A, 99 boys, 79 girls; General High School B, 163 boys, 127 girls; General High School C, 95 boys, 90 girls; Technical High School A, 71 boys, 56 girls; *single-sex schools*—General High School D, 143 boys; Technical High School B, 153 boys; General High School E, 126 girls; and Technical

[3] It is obviously very difficult to compare the results of the two studies just described. One would have to make allowance for factors associated with the different cultures, for differences in the assessment procedures that were employed, for differences in subject characteristics, and indeed for a host of varying factors that would affect comparisons.

High School C, 105 girls. Most children were in the age range 15–17 years. The youngest boys and the youngest girls were those in the single-sex technical high schools, where the proportion of final year matriculating students was very low. Status of father's occupation, as measured by the Congalton Index (Congalton, 1969) was slightly higher for both boys and girls in the coeducational high schools than for boys and girls in the single-sex schools and slightly higher for girls only in the general high schools when compared to girls in the technical high schools.

Procedure

Details of test administration, including information about the tests used and specific instructions, have appeared elsewhere (Feather, 1972a). On Form E of the Rokeach Value Survey subjects are normally required to rank sets of values (18 *terminal* values and 18 *instrumental* values) in their order of importance, the values being presented alphabetically and with short descriptive phrases. In the present survey there was not sufficient time to permit subjects to rank both sets of values; the questionnaires were randomly distributed so that approximately half the subjects ranked the terminal values only, while the remaining subjects ranked the instrumental values only. Each subject ranked whichever set of values he or she had in two ways, first in relation to self, then in relation to the school he or she attended. The first set of instructions asked subjects to study the list of 18 values carefully, then to place a 1 next to the value "which is most important to you, place a 2 next to the value which is second most important to you, etc. The value which is least important, relative to the others, should be ranked 18." The second set of instructions asked subjects to "assume that your School is attempting to turn out children with certain kinds of values, some of which the School considers to be more important than others." They were then asked to "think of a student who has these values that the School strives to emphasize and think of the order in which the School would emphasize them." They then ranked the values 1–18 according to how they thought the school would emphasize them from most important to least important. The two sets of rankings provided for self and then for school represented *own value systems* and *school value systems*, respectively.

A modified form of the Cornell Job Description Index (Smith, 1967; Smith et al., 1969) developed by the author to apply to the school situation was used to obtain measures of school satisfaction. This modified Job Description Index required subjects to check lists of 18 items as they applied to schoolwork, people in my class, and the typical teacher. Separate satisfaction scores that could range 0–18 in the direction of increasing satisfaction were obtained for each of these three aspects of the school situation (see Feather, 1972a).

Finally, subjects were asked to rate how much they enjoyed being at school by putting a cross on a five-inch scale labeled "Very happy at school" at one extreme, "Don't like school at all" at the other extreme, and "Moderately happy at school" in the middle of the scale. Responses were scored 1–9 in the direction of increasing happiness, with a score of 5 spanning the midpoint of the scale.

Results and Discussion

Analysis of Average Value Systems

The median rankings for each set of 18 values (terminal or instrumental) were calculated for each of the eight schools for both self (own) values and for school values. For the set of 18 terminal values, there were 16 different sets of medians (or average value systems) corresponding to the combination of eight schools and two types of ranking (own values and school values). These 16 average orders were intercorrelated using the Spearman rank-order procedure to yield a 16 × 16 matrix for the terminal values. This correlation matrix, with unities in the diagonal cells, was factor analyzed using the principal-component method in conjunction with varimax rotation for eigenvalues greater than one. Two factors emerged; the first, accounting for 47.32% of the variance, could be clearly identified as a school value factor, indicating a basic similarity in the median rankings of school terminal values across all conditions. The second factor, accounting for 42.57% of the variance, could be clearly identified as an own value factor, indicating a basic similarity in the median rankings of own terminal values across all conditions. The nature of these average orders is indicated in Table 1 for boys and girls in coeducational and single-sex schools.

The same procedure was applied to the 16 average orders for the set of 18 instrumental values. A school value factor also emerged as the first factor in this analysis, accounting for 48.33% of the variance. Two other factors were also extracted, both clearly identifiable. The first, accounting for 22.86% of the variance, was a female own value factor, indicating a basic similarity in the median rankings of own instrumental values across females. The second, accounting for 20.21% of the variance, was a male

TABLE 1

MEDIAN RANKINGS OF TERMINAL VALUES IN RELATION TO SELF AND SCHOOL FOR BOYS AND GIRLS IN COEDUCATIONAL AND SINGLE-SEX SCHOOLS

Terminal value	Boys				Girls			
	Coeducational		Single sex		Coeducational		Single sex	
	Own	School	Own	School	Own	School	Own	School
A comfortable life	10.79	8.82	9.13	7.78	13.96	10.29	13.56	11.43
An exciting life	7.56	11.58	8.83	12.63	9.88	11.80	9.57	11.75
A sense of accomplishment	7.50	2.53	8.06	2.44	7.73	1.74	8.57	1.85
A world at peace	5.33	8.38	5.44	8.25	4.77	8.11	3.60	7.17
A world of beauty	12.82	11.81	12.05	10.82	11.53	10.93	12.00	10.38
Equality	7.68	5.68	7.14	5.46	5.59	5.36	6.33	5.40
Family security	8.38	10.50	6.50	9.83	7.21	8.86	7.50	9.29
Freedom	4.57	7.50	6.19	7.50	5.55	7.27	5.57	8.71
Happiness	7.08	8.88	6.94	9.60	6.46	10.11	6.43	10.14
Inner harmony	9.95	11.57	11.75	10.69	8.25	10.86	8.20	10.80
Mature love	6.43	15.24	6.32	14.75	7.89	16.06	8.27	16.41
National security	13.31	10.83	14.42	9.00	13.59	9.60	12.59	10.33
Pleasure	11.03	11.79	10.06	12.00	13.03	13.57	12.71	13.00
Salvation	16.09	14.29	16.19	13.82	15.04	15.00	15.00	12.50
Self-respect	10.60	5.57	10.46	6.54	9.75	6.00	10.64	5.77
Social recognition	14.40	6.59	13.54	6.14	14.09	6.31	14.15	4.90
True friendship	5.38	7.74	4.97	9.00	5.18	8.73	4.70	8.42
Wisdom	7.50	2.54	7.50	3.44	6.05	2.00	6.45	2.53

Note. In each column low numbers denote high relative value.

own value factor, indicating a basic similarity in the median rankings of own instrumental values across males. The nature of these average orders is indicated in Table 2 for boys and girls in coeducational and single-sex schools.

Thus, as in other studies (Feather, 1972b, 1973), factors involving own and school values emerged. There were basic disparities in the way students ranked their own values and in the way they ranked the values for their school, but neither factor analysis provided evidence for a factor contrasting coeducational high schools with single-sex high schools.

Analysis of Particular Values

The rankings for each value were separately analyzed using 2 × 2 analyses of variance involving Sex Composition of School (coed vs. single sex) × Type of School (general vs. technical) as factors and using the method of unweighted means to allow for unequal *n*s (Winer, 1962). As in previous studies (Feather, 1972a), the rankings were first transformed using the normal curve (Hays, 1967, pp. 35–39). The analyses were run for males and females separately, first for the transformed rankings of own values then for the transformed rankings of school values. A conservative alpha level ($p < .01$) was adopted to test significance in view of the ipsative nature of the ranking procedure and the large number of comparisons that were made.

There were no significant differences between the coeducational and single-sex schools in the relative importance assigned to particular *terminal* values by either boys or girls for either self or school rankings. In regard to particular *instrumental* values, however, both boys and girls from single-sex schools ranked being clean as higher in relative importance than did students from coeducational schools; this was true for both self and school rankings. Girls in single-sex schools also saw their schools as placing relatively more emphasis upon being helpful and polite than did girls in coeducational schools; they also saw their schools as as-

TABLE 2

MEDIAN RANKINGS OF INSTRUMENTAL VALUES IN RELATION TO SELF AND SCHOOL FOR BOYS AND GIRLS IN COEDUCATIONAL AND SINGLE-SEX SCHOOLS

Instrumental value	Boys				Girls			
	Coeducational		Single sex		Coeducational		Single sex	
	Own	School	Own	School	Own	School	Own	School
Ambitious	7.40	2.81	6.20	2.85	9.41	1.76	8.75	2.00
Broadminded	6.77	11.11	7.25	10.04	5.96	12.06	7.30	12.11
Capable	8.21	5.98	8.36	6.63	9.41	6.75	9.81	6.25
Cheerful	8.75	13.50	9.40	14.61	7.54	13.74	6.94	13.05
Clean	10.53	8.08	8.38	6.65	9.81	8.54	7.75	6.58
Courageous	9.90	13.75	9.56	14.03	9.83	13.61	11.17	14.05
Forgiving	10.08	14.14	11.15	14.46	8.65	14.37	8.36	14.15
Helpful	10.38	10.54	12.00	10.56	7.88	10.89	8.65	7.70
Honest	4.70	5.82	3.90	5.93	2.58	6.11	3.25	5.75
Imaginative	13.06	11.68	13.33	10.06	14.09	12.35	15.05	13.15
Independent	9.00	11.26	7.43	10.69	11.58	10.50	10.59	11.44
Intellectual	11.46	5.21	10.71	6.90	14.58	5.50	12.95	7.83
Logical	8.91	8.31	10.08	7.83	12.21	8.27	12.79	9.70
Loving	10.09	17.06	10.00	16.45	7.83	17.23	7.19	17.07
Obedient	13.00	4.56	12.11	5.61	11.54	4.16	10.69	5.33
Polite	9.40	5.56	9.83	6.81	8.16	5.09	7.28	3.70
Responsible	5.00	4.48	4.63	5.03	3.80	4.33	4.93	4.36
Self-controlled	7.88	8.09	7.55	8.68	8.58	7.31	9.79	7.67

Note. In each column low numbers denote high relative value.

signing less importance to being intellectual and logical.

There was only one significant effect involving the Sex Composition of School × Type of School interaction (at the conservative alpha level, $p < .01$): Boys in the coeducational technical high school ranked being forgiving as relatively more important for themselves than did boys in the single-sex technical high school, but there was no such difference among boys in the general high schools.[4]

In summary, therefore, main effects of sex composition of school (coed vs. single sex) were found only in regard to the instrumental values and then, with the exception of being clean, only for the girls and only for the perceived school values. Some of the differences supported Dale's suggestion that single-sex schools might assign more importance to rules of conduct, but there were other values related to control and conduct (e.g., being obedient) that showed no significant differences at all. Nor was there any support for Coleman's suggestion that coed students might set greater store upon being popular and accepted by their peers than would students in the single-sex secondary schools.

There were some main effects of type of school (general vs. technical) on value rankings. In particular, for both boys and girls, inner harmony was ranked as relatively less important and a comfortable life as relatively more important in the technical high schools than in the general high schools when rankings were for self. These results may reflect the greater vocational emphasis in the technical high schools.

[4] Among the terminal values, boys ranked the following values as relatively more important for self than did girls: a comfortable life, mature love, and pleasure. But girls ranked inner harmony as relatively more important for self than did boys. Among the instrumental values, boys ranked the following values as relatively more important for self than did girls: being ambitious, imaginative, independent, intellectual, and logical. But girls ranked the following values as relatively more important for self than did boys: being helpful, honest, loving, and polite. All of these differences between boys and girls were highly significant ($p < .01$ or $p < .001$).

TABLE 3

MEAN SATISFACTION AND HAPPINESS SCORES FOR BOYS AND GIRLS IN RELATION TO SCHOOL ATTENDED

Type and composition of school	*M* satisfaction scores			*M* happiness rating
	Schoolwork	People in my class	Typical teacher	
Boys				
Coeducational				
General	7.81	10.22	10.90	5.08
Technical	8.56	10.57	11.55	5.64
Single sex				
General	6.32	8.92	10.53	4.84
Technical	8.86	9.55	10.63	5.23
Girls				
Coeducational				
General	7.40	11.36	10.19	5.22
Technical	8.48	10.04	11.43	5.73
Single sex				
General	8.24	10.58	10.06	5.34
Technical	8.45	11.43	11.99	5.29

Analysis of Satisfaction with School

Table 3 presents the mean satisfaction scores and mean happiness ratings for boys and girls in relation to school attended. The 2 × 2 analyses of variance applied to the data for the boys indicated that boys in the coeducational high schools reported more satisfaction both with people in my class and with the typical teacher than did boys in the single-sex high schools ($p < .001$ and $p < .05$, respectively). There was a significant Sex Composition of School × Type of School interaction in regard to satisfaction with schoolwork ($p < .01$): Boys in the coeducational general high schools reported more satisfaction with their schoolwork than did boys in the single-sex general high school, but this difference was in the reverse direction for boys in the technical high schools, although only marginally so.

The only significant effect involving coeducation in the 2 × 2 analyses of the data for the girls was a Sex Composition of School × Type of School interaction ($p < .01$): Girls in the coeducational general high schools expressed greater satisfaction with people in my class than did girls in the single-sex general high school, but this difference was reversed for girls in the technical high schools. There, the girls in the single-sex school expressed greater satisfaction for people in my class.

The girls in general tended to report greater satisfaction with people in my class than did the boys ($p < .001$), regardless of whether their school was coeducational or single-sex or general or technical.

There were some significant main effects of type of school (general vs. technical) in relation to the measures of satisfaction. Both boys and girls in the technical high schools expressed greater satisfaction with schoolwork than did boys and girls in the general high schools ($p < .001$ for boys; $p < .05$ for girls). Girls in the technical high schools reported greater satisfaction with the typical teacher than did girls in the general high schools ($p < .001$). Boys in the technical high schools expressed greater happiness with school than did boys in the general high schools ($p < .05$). In the technical high schools, children may see their schoolwork as more relevant to future careers than do children in the general high schools, where the vocational emphasis tends to be much less apparent and where the curriculum is rather more academic and general.

CONCLUSIONS

The results of the present study indicated that there were few differences in the way boys and girls from coeducational and single-sex schools assigned importance to values. The evidence was somewhat stronger for differences in attitudes toward aspects of the school situation when children from coeducational schools were compared to children from single-sex schools. There was no support for Coleman's hypothesis concerning possible adverse effects of coeducation and limited support for Dale's suggestion that single-sex schools may be seen as more concerned with discipline and control. As far as values were concerned, the major distinction was between the value systems the students assigned to themselves and the order of values they believed their schools were trying to promote.

In future investigations, one might em-

ploy finer grain measures in the hope that they would be more responsive to the effects of coeducation and, if possible, take a much wider sample of different types of school, both single-sex and coeducational. In the foreseeable future, however, finding a single-sex school within the state school system might (as one headmaster put it) be like looking for a saber-toothed tiger. Moreover, the move toward coeducation in Australia is a general one now beginning to extend to some of the independent schools as well. Hence the present study is in the unique position of presenting information about a division that is rapidly disappearing.

REFERENCES

COLEMAN, J. S. *The adolescent society.* New York: Free Press of Glencoe, 1961.

CONGALTON, A. A. *Status and prestige in Australia.* Melbourne, Australia: Cheshire, 1969.

DALE, R. R. Co-education—I: A critical analysis of research on the effects of co-education on academic attainment in grammar schools. *Educational Research,* 1962, **4,** 207–217. (a)

DALE, R. R. Co-education—II: An analysis of research on comparative attainment in mathematics in single-sex and co-educational maintained grammar schools. *Educational Research,* 1962, **5,** 10–15. (b)

DALE, R. R. Co-education—III: Research on comparative attainment in English in single-sex and co-educational grammar schools. *Educational Research,* 1964, **6,** 170–178.

DALE, R. R. *Mixed or single-sex school?* Vol. I. London: Routledge & Kegan Paul, 1969.

DALE, R. R. *Mixed or single-sex school?* Vol. II. London: Routledge & Kegan Paul, 1971.

FEATHER, N. T. Value similarity and school adjustment. *Australian Journal of Psychology,* 1972, **24,** 193–288. (a)

FEATHER, N. T. Value similarity and value systems in state and independent secondary schools. *Australian Journal of Psychology,* 1972, **24,** 305–315. (b)

FEATHER, N. T. Value systems and education: The Flinders programme of value research. *Australian Journal of Education,* 1972, **16,** 136–149. (c)

FEATHER, N. T. Effects of response anonymity on assessment of own and school value systems and satisfaction with school. *British Journal of Educational Psychology,* 1973, **43,** 140–150.

HAYES, W. L. *Quantification in psychology.* Belmont, Calif.: Brooks/Cole, 1967.

JONES, J. C., SHALLCRASS, J., & DENNIS, C. C. Coeducation and adolescent values. *Journal of Educational Psychology,* 1972, **63,** 334–341.

ROKEACH, M. *Beliefs, attitudes, and values.* San Francisco: Jossey-Bass, 1968. (a)

ROKEACH, M. A theory of organization and change in value-attitude systems. *Journal of Social Issues,* 1968, **24,** 13–33. (b)

ROKEACH, M. The measurement of values and value systems. In G. Abcarian & J. W. Soule (Eds.), *Social psychology and political behavior.* Columbus, O.: Merrill, 1971.

ROKEACH, M. *The nature of human values.* Riverside, N.J.: Free Press, 1973.

SMITH, P. C. The development of a method of measuring job satisfaction: The Cornell studies. In E. A. Fleishman (Ed.), *Studies in personnel and industrial psychology.* Homewood, Ill.: Dorsey, 1967.

SMITH, P. C., KENDALL, L. M., & HULIN, C. I. *The measurement of satisfaction in work and retirement.* Chicago: Rand McNally, 1969.

WILLIAMS, R. J., JR. Change and stability in values and value systems. In B. Barber & A. Inkeles (Eds.), *Stability and social change.* Boston: Little, Brown, 1971.

WINER, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received January 3, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 16-28

REMEDIATION OF LEARNING PROBLEMS AMONG THE DISADVANTAGED[1]

SEYMOUR FESHBACH[2] AND HOWARD ADELMAN

University of California, Los Angeles

In an effort to help clarify whether variations in school input can remedy academic deficiencies among the disadvantaged, a three-year experimental project was carried out. Specifically, the focus was on the impact of an intensive, individualized, and integrated remedial program on a group of disadvantaged youngsters with learning problems. The major findings are that (*a*) The disadvantaged students learned as effectively as a comparable group of middle-class youngsters and significantly better than disadvantaged youngsters enrolled in a more traditional compensatory education program, and (*b*) these findings held for young adolescents as well as elementary-age children. The study is seen as providing evidence of the favorable consequences for the disadvantaged of *comprehensive,* as contrasted with piecemeal, educational programs.

There are few issues in American education that are the object of as much attention and controversy as the educational problems presented by so-called "disadvantaged" [3] populations (Ornstein, Doll, Arnez, & Hawkins, 1971). Indeed, the significance of these problems extends beyond education, having profound social, economic, and political implications as well.

There is little disagreement as to the descriptive aspects of the problem, namely, that children from economically disadvantaged populations perform more poorly on measures of academic achievement than children from advantaged populations. Sharply differing explanations, however, have been offered to account for such findings. In the case of blacks, a significant proportion of whom can be categorized as disadvantaged, one view suggests that racially linked genetic factors are responsible for race differences in IQ and academic achievement (Jensen, 1969). A more widely held explanatory view focuses on pupil attitudinal and cognitive characteristics that arise from variations in home environment (Deutsch, 1968; Hess, Shipman, Brophy, & Bear, 1968; Hunt, 1968). Both genetic and early environmental social pathology explanations have been pointedly criticized by Baratz and Baratz (1970), who suggest that the principal aetiological factor is the inappropriateness of most school programs for the black disadvantaged child. A number of other writers have also stressed the inadequacy of the typical educational structure and program to which the disad-

---

[1] The project was supported by funds from the California State Department of Education, Division of Compensatory Education through the Bureau of Professional Development's Research and Teacher Education (RATE) program.

The authors wish to express a debt of gratitude to the many individuals at the Fernald School and in the Los Angeles Public Schools who participated in the various aspects of this project. We wish to particularly acknowledge the help of Frances Berres, Associate Head of the Fernald School, who assumed a major responsibility in administering the Fernald School phase of the program. Special mention should also be made of the devotion and assistance given to various phases of the project by John Long and Williamson W. Fuller III.

[2] Requests for reprints should be sent to Seymour Feshbach, Department of Psychology, University of California, Los Angeles, California 90024.

[3] As defined in Title I of the Elementary and Secondary Education Act of 1965, the term "disadvantaged" designates pupils from families with income below $3,000 per year. It is clear, however, that the term also is used to designate segments of racial or ethnic minority groups and often is intended to connote that such groups are culturally different.

vantaged child is subject, although not necessarily agreeing on which school characteristics are ineffective or counterproductive (Clark, 1965; Fantini & Weinstein, 1968). A related, alternative position focuses on the interaction between pupil and school characteristics (Deutsch, 1963; Feshbach & Adelman, 1971).

These diverse views of the antecedents of academic deficiencies among the disadvantaged have led to numerous variations in school inputs, but because of the apparent lack of success of remedial and enrichment programs at most age levels, there is a general pessimism regarding such programs. Nevertheless, it remains an open question as to whether variations in school input can result in the remediation of academic deficiencies among the disadvantaged. The primary objective of the present investigation was to contribute data that could help resolve this issue. Specifically, the focus was on the impact of an intensive, individualized, and integrated remedial program upon academic achievement, intellectual and perceptual–cognitive functioning, aspiration levels, and self-attitudes of a group of disadvantaged youngsters who manifested school problems. For purposes of comparison, the impact of this intensive program was contrasted with a more traditional compensatory education program. In addition, the study provided the opportunity to explore similarities and differences between groups of disadvantaged and middle-class children, all of whom manifested learning problems, with reference to a number of the variables under investigation.

Method

Subjects

Both the disadvantaged and advantaged students who participated in the project met the following criteria. They all were (*a*) male, (*b*) of at least average intelligence (in a few instances, a youngster with an IQ in the high 80s was included if the data in his records suggested that the IQ indicated might underestimate his true ability), (*c*) one and one-half or more years retarded in basic school skills, and (*d*) without severe neurological or severe emotional disturbances.

In addition, the disadvantaged students lived in areas that were designated as poverty pockets, that is, average family income was approximately $3,000 a year. (It is recognized that the economic criterion does not adequately define the concept of disadvantaged as this term has come to be used in current literature. Nevertheless, income is undoubtedly the best single criterion and predictor of a disadvantaged condition.) These children were chosen from a list that the counselors at each participating school prepared to conform with the above criteria. From these lists, the project staff selected different children for participation in the project during the academic years 1966–67, 1967–68, and 1968–69. During the first academic year, 30 elementary and 30 junior high disadvantaged youngsters participated in the study; during the subsequent two years, 50 elementary and 30 junior high disadvantaged youngsters participated. In each of these years, approximately 90% of the subjects were black.

The advantaged students were all selected from the tuition-paying clients enrolled at the Fernald School. This school, a facility of the psychology department of the University of California, Los Angeles, is a research and training laboratory focusing on learning disorders not due to mental retardation or severe neurological or emotional pathology. With few exceptions, the advantaged students were middle- or upper-class Anglos.

Design

During the first year, the 60 disadvantaged youngsters selected were grouped into triplets, matched for age, IQ, race, and severity of learning deficit. From each triplet, one student was randomly assigned to the group that was bussed to the Fernald School; another student was assigned to the school enrichment program, which was conducted in the home schools; and the third was assigned to a control group. In this way, each group was assigned 10 elementary and 10 junior high students. Then, a group of 10 elementary and 10 junior high advantaged youngsters was identified from the regular Fernald School population to form a fourth, comparison group, matched for age, learning disability, and approximate IQ with the disadvantaged samples.

In the second year of the program, although there were several changes, this same general design and procedure was followed. The changes involved (*a*) doubling the number of elementary students in the school enrichment and control groups (from 10 to 20), (*b*) including three rather than two public schools, and (*c*) selecting public schools in a different section of Los Angeles, that is, in the first year, the students had come from the Venice area on the western boundary of the city, and in the second year they came from mid-city.

Remedial Programs

Two remedial programs were implemented and evaluated during this study: (*a*) the Fernald School program and (*b*) the school enrichment program. The Fernald School program differs from standard

school experiences in (*a*) the high degree to which the classroom program is individualized and extra classroom supports are available, (*b*) the relatively low pupil-to-teacher ratio, and (*c*) the "special" qualities of the total environment, for example, an atmosphere of experimentation, emphasis on positive reinforcement, and the reduction of school-related anxieties. An extensive discussion of the rationale underlying this program and of its characteristics is presented elsewhere (Fernald School, 1969).

The school enrichment program represented an attempted to provide a traditional, but high quality, type of compensatory intervention program. The individualized remedial procedures were patterned after those employed at the Fernald School, but, of course, other features of the Fernald program could not be exported, and the time devoted to remediation had to be limited. Consequently, the primary focus was on improving reading and language skills during three to five hours per week.

In the first year, the program essentially supplemented the regular teachers' reading lessons. Each of three Fernald teachers took a group of three to four pupils for one hour per day, three days per week. However, since a goal of the project was to have a wider impact, during the second and third years Fernald School teachers took total charge of the reading programs of the participating pupils. That is, the teachers worked simultaneously with the entire school enrichment group from a particular school. It was felt at the time that such an approach might result in a youngster's regular classroom teacher discussing other facets of his program with the Fernald teachers, thereby allowing them to present ideas for introducing individualized instruction into the regular classroom programs. (With reference to both remedial programs, it was recognized that pupils could well have benefited from participating for more than one year. Unfortunately, resource limitations prevented this course of action.)

Measures

Various combinations of measures were used during each phase of the project. The choice of measures was guided by our interest in achievement gains and related intellectual and perceptual–cognitive functioning. Since no completely satisfactory instruments were available with reference to the variables that were of interest, we tended to choose instruments that had been found useful by other investigators. The instruments relevant to the present discussion, used one or more times during the three years of the project, are:

1. California Achievement Test (Tiegs & Clark, 1957, with norms revised in 1963);
2. Wechsler Intelligence Scale for Children (based on the work of Enburg, Rowley, and Stone [1961] and Carleton and Stacey [1954], a short form including six of the ten standard subtests was used);
3. Auditory Discrimination Test (Wepman, 1958);
4. Visual Motor Gestalt Test (Bender, 1946);
5. Frostig Developmental Test of Visual Perception (Frostig, Lefever, & Whittlesey, 1966);
6. Vocational Checklist—boys' form (Wrightstone, Forlano, Frankel, Lewis, Turner, & Bolger, 1964);
7. Ethnic Attitudes Instrument (Gerard, 1969);
8. Locus of Control Scale (James, 1957).

It should be noted that due to administrative policies in the Los Angeles City School District, some variations in procedure were required in administering these measures to the school enrichment and control groups, and, indeed, some instruments could not be given at all to these two groups. Other reasons for changes in the assessment procedures were (*a*) if a measure proved to be unreliable and to have limited utility and (*b*) if certain supplementary studies required the addition of particular measures.

Results and Discussion

At the onset it was emphasized that while we have relied heavily on quantitative procedures, our available measuring instruments could only capture a restricted segment of the behavior being assessed. Reading achievement tests reflect only a part of a child's achievement in reading; our measure of vocational aspiration only tapped the surface of the child's feelings about his vocational options and likely future. And there are domains of behavior that were not assessed at all, but that both teacher and pupil may perceive as significant. Clearly, the issues are complex; the variables involved are many; the methodology is imperfect.

To partially compensate for the dryness and, more particularly, for the limitations of our quantitative analyses, we have incorporated a number of qualitative observations in this report. While the quantitative analyses, albeit in a limited way, can stand by themselves, the qualitative observations cannot, and therefore, are interpreted in conjunction with the numerical findings.

Age and IQ Characteristics of the Samples

Before reviewing the major experimental findings, it is helpful to consider some of the characteristics of our experimental samples. The number of subjects in each experimental group who participated in at least one pre–post measure is presented by school year and for all three years in Table 1. As is almost inevitable in a field study of this

TABLE 1
NUMBER OF SUBJECTS IN EACH EXPERIMENTAL GROUP WHO PARTICIPATED IN AT LEAST ONE PRE-POST MEASURE

Level	Fernald advantaged	Fernald disadvantaged	School enrichment	Control
School year 1966–67				
Elementary	13	9	9	7
Junior high	11	8	10	10
School year 1967–68				
Elementary	9	9	15	16
Junior high	11	11	9	9
School year 1968–69				
Elementary	10	10	18	16
Junior high	9	9	10	9
All three school years combined				
Elementary	32	28	42	39
Junior high	31	28	29	28

kind, the number of subjects who completed the study differs from the number initially selected.

In the advantaged group, there were actually a few more subjects than the number that had been planned. These were added, largely in the first year, because they were available and also improved the matching.

The mean ages and IQs of the subjects included in each experimental group are presented in Table 2. The IQs of the advantaged are higher than those of the disadvantaged boys. This difference was anticipated, since a more liberal criterion was used in selecting disadvantaged children with at least "average" IQs.

The basis for the differences in IQ among the disadvantaged elementary children is less clear. These children were randomly assigned to the various experimental conditions and should have comparable IQ scores. One minor factor contributing to the reliably lower IQ of the disadvantaged elementary group is the somewhat greater mean IQ of the children in the enrichment and control groups who remained in the project as compared to the children who were not available for posttesting. Another possible source of bias lies in the initial selection. While the children were randomly assigned to each group and while the great majority of families then agreed to send their child to the Fernald School, some substitutions may have been lower IQ children. However, as the subsequent analysis shows, the experimental effects found for the elementary age children were similar to those obtained for the junior high school subjects where there were no significant group differences in IQ.

The initial IQ differences for the elementary subjects are reflected in the pretest measures of achievement. The initial California Achievement Test mean grade placement scores of the Fernald Elementary disadvantaged children was 2.71; the mean for the controls was 3.32; while the mean for

TABLE 2
MEANS AND STANDARD DEVIATIONS OF AGE AND IQ OF EXPERIMENTAL GROUPS

Level	Advantaged		Disadvantaged		Enrichment		Control	
	IQ	Age	IQ	Age	IQ	Age	IQ	Age
M (in months for age)								
Elementary	98.2	118.7	91.5	117.4	94.3	115.5	96.5	116.9
n	32	32	28	28	41	42	39	39
Junior high	97.7	159.9	91.3	158.9	93.3	157.6	91.3	158.1
n	31	31	27	28	28	29	26	28
SD								
Elementary	8.6	12.4	8.7	13.2	10.7	13.3	9.7	13.9
Junior high	6.9	11.9	7.5	7.0	8.6	8.2	8.8	7.9

TABLE 3
CALIFORNIA ACHIEVEMENT TEST TOTAL GRADE PLACEMENT

Level	Measure	Fernald advantaged	Fernald disadvantaged	School enrichment	Control
		Pretest *M*			
Elementary	*M*	2.75	2.71	2.87	3.32
	SD	.99	.90	1.28	.79
	n	32	28	39	36
Junior high	*M*	5.98	6.15	5.76	5.50
	SD	1.06	1.11	1.34	1.56
	n	28	28	28	27
Total	*M*	4.26	4.43	4.08	4.25
	SD	1.92	2.00	1.98	1.60
	n	60	56	67	63
		Change score *M*			
Elementary	*M*	1.08	1.06	.68	.75
	SD	.65	.49	.73	.47
	n	32	28	39	36
Junior high	*M*	1.04	1.10	.57	.52
	SD	.64	.46	.47	.57
	n	28	28	28	27
Total	*M*	1.06	1.08	.63	.65
	SD	.64	.47	.63	.52
	n	60	56	67	63

the controls who left the project was 2.96. While pretest differences were present at the elementary level, these did not materially influence the outcome of the study. Thus, there was a negligible relationship between initial level and amount of change (a rather surprising finding in view of the statistical regression effects). Second, comparable effects were observed at the junior high level. Finally, special analyses were undertaken in which the effects of initial differences in pretest scores on subsequent posttest scores were eliminated through statistical procedures (covariance), and these analyses yielded results that were practically identical to the comparisons of the amount of change displayed by each group.[4]

[4] These and the other statistical references are two-way (Condition × Age Group) analyses of variance and covariance performed using Biomedical Computer Program BMD ×64 "General Linear Hypothesis," written by Paul Sampson of the Health Sciences Computing Facility, University of California, Los Angeles. Specific comparisons among experimental groups were made as subanalyses within the overall analyses. For further information on these procedures see Dixon (1969), Kempthorne et al. (1961), and Scheffé (1959).

Group Changes and Differences

Achievement testing. The first and most important empirical question to be considered concerns the degree of movement in basic academic skills in the disadvantaged children who participated in the intensive, individualized, and socially integrated Fernald School program. The data on these children are contrasted with the academic progress of (*a*) the advantaged children enrolled in the Fernald School program, (*b*) the disadvantaged children who participated in the more traditional compensatory (school enrichment) program, and (*c*) the control group of disadvantaged youngsters.

The California Achievement Test was administered at the beginning and at the end of each academic year to all groups. This test consists of three components, reading, arithmetic, and language skills (each of which has two subscales), and yields a total score descriptive of the child's overall grade placement equivalent. The pretest and change score means of these California Achievement Test total grade placement scores are presented in Table 3.

EDUCATIONAL BUREAU OF LIBRARY CALCUTTA-19

Reference has been already made to the IQ differences between the Fernald disadvantaged and control groups at the elementary level, and the statistically reliable difference in initial grade placement scores is probably a reflection of this IQ difference. The mean IQs of the Fernald disadvantaged and control students are comparable at the junior high level; therefore, the differences in initial grade placement scores at the junior high level (also statistically reliable) cannot be attributed to IQ differences.

The use of the IQ enables us to predict at a much better than chance level a child's response in different learning situations. At the same time, we know that children with the same IQ manifest markedly different learning patterns and that the same child's performance varies greatly in different learning circumstances. From the beginning of our work with the disadvantaged children, we were struck with their responsiveness to different school and class situations and the variation in their behavior under these different conditions. During the first day of testing, the children tested at their home schools were restless, defensive, nonconforming, and negativistic. The matched group of children tested at the Fernald School, particularly the junior high boys, were obliging, serious, and task oriented. These behavioral differences may well have influenced the achievement test performance.

While the pretest differences are of interest and importance, the key data lie in the change score means presented in Table 3. The movement of the two groups of children at the Fernald School is remarkably similar. At both the elementary and junior high level and for both advantaged and disadvantaged samples, the increase in grade placement is about a year and a month. In contrast, the movement in both the enrichment and control groups was significantly less. The increase in the Fernald disadvantaged group is significantly greater than the corresponding increase in the control and in the enrichment groups at both the elementary and junior high levels (based on both analyses of variance and covariance). At the junior high level, the increase of 1.10 in the children who attended the Fernald School is about twice the amount of change in the school enrichment and control groups. The findings indicating that junior high level children can derive substantial benefits from a remedial program merits special attention in view of the widely held opinion that compensatory remedial educational efforts are relatively ineffective for this population and are best expended at earlier ages.

These data then indicate the following:

1. The disadvantaged children who attended the Fernald School and the advantaged children at the Fernald School made increases in grade placement scores of slightly more than one year.

2. The disadvantaged children who attended the Fernald School made significantly greater gains than either the enrichment or control groups.

3. The relative advantage of the Fernald disadvantaged children over the other groups was most pronounced at the junior high level.

4. The enrichment children did *not* make significantly greater gains than did the controls.

This pattern of findings holds for many of the subtests of the overall achievement test scale. However, the differences in the disadvantaged groups in reading achievement were largely due to the great gains made by the Fernald children in reading comprehension. (The children's teachers felt that the Fernald and enrichment groups did make substantial gains in reading vocabulary, but that these gains were not reflected in the California Achievement Test measure, which tends to sample middle-class rather than lower-class linguistic terms. This possible bias could be particularly acute because the individualized methods used emphasize, as one important source of new reading vocabulary, the concepts that the child employs in his speech and in his story writing.)

With reference to arithmetic achievement totals, there was very little difference at the elementary level between the Fernald disadvantaged group and the other two groups in changes on the arithmetic reasoning subtest, while the differences in arithme-

Edu. & Psyl. Research (N.C.E.R.T.) T 886

tic fundamentals were larger and consistent with the overall trend. At the junior high level, the gain in arithmetic fundamentals in the enrichment and control samples was negligible and significantly smaller than that of the Fernald disadvantaged group who showed a year's increment. The change in arithmetic reasoning in the Fernald disadvantaged junior high group was particularly impressive. This group's mean gain of 1.3 years was significantly greater than the gain of .9 years in the advantaged group and .7 and .6 years in the enrichment and control groups, respectively. Since the skills entailed in arithmetic reasoning are of a higher conceptual order than the more rote content of arithmetic fundamentals, the gain achieved by the Fernald disadvantaged children acquires special significance. (It should be reemphasized that the enrichment group was not given special instruction in the arithmetic area.)

While the differences between the school enrichment and control groups on the total language scale scores were slight, the disadvantaged children at the Fernald School attained the highest scores of all the groups, significantly different from those of the enrichment groups and from those of the controls, although not from those of the advantaged children. Again, the differences at the junior high level were larger than those at the elementary level. For the younger children, the difference between the Fernald disadvantaged students and the controls was statistically reliable, but that between the Fernald and the enrichment group failed to achieve statistical significance. While the control group made the least gain in spelling, the increments in the other groups were not much larger. None of the differences were statistically reliable, and they contributed in only a minor way to the differences obtained on the total language measure. The main source of these differences was the large increment obtained by the Fernald disadvantaged children on the English mechanics subtest. At the elementary level, the Fernald disadvantaged children increased a little more than 1.1 years, and at the junior high level, they made a gain of 1.4 years in English mechanics. The latter gain was significantly greater than that achieved by either the enrichment or control groups. At the elementary level, these differences only attain the .10 level of significance.

Intelligence testing. Six subtests of the Wechsler Intelligence Scale for Children (WISC) comprehension, vocabulary, arithmetic, similarities, picture arrangement, and block design, were administered at the beginning of the year, and due to limited availability of testing time, especially for the enrichment and control groups, the first three of these subtests were given again at the end of the academic year only for the second and third years of the project. The numbers of boys in each experimental group used in analysis of these data are therefore fewer than the numbers available for the achievement comparisons, and reliable results are more difficult to obtain. The pretest and change means for the three repeated subtests are presented in Table 4.

While the differences in comprehension pretest means are not statistically reliable, the higher initial score of the control elementary group is consistent with the higher initial scores attained on the achievement measures. The change scores on this subtest are quite variable, and none of the differences are statistically significant. As an incidental note, the fact that the elementary and junior high groups have comparable scores should not be interpreted to mean that their absolute performance was the same. The numbers in the pretest table are weighted scores based on the age of the child as well as his performance. (A weighted score of 10 on each subtest is equivalent to an IQ of 100; an average of 7 on each is equal to an IQ of 94; and an average of 8 is 96. The standard deviation of each scale is 3.)

The vocabulary scores on the pretest of the advantaged children are significantly higher than those of the disadvantaged groups, while the differences among the latter are not statistically reliable. The superiority of the advantaged children on the vocabulary subscale of the WISC is consistent with the results of other studies comparing the linguistic repertoire of advantaged and

TABLE 4

WECHSLER INTELLIGENCE SCALE FOR CHILDREN: COMPREHENSION, VOCABULARY, AND ARITHMETIC SUBTESTS

Test	Fernald advantaged			Fernald disadvantaged			School enrichment			Control		
	M	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Pretest *M*												
Comprehension												
Elementary	7.89	2.79	19	8.42	1.39	19	8.81	2.96	32	9.84	2.69	32
Junior high	9.22	2.16	18	8.06	1.73	18	8.78	3.10	18	7.94	2.21	18
Vocabulary												
Elementary	10.67	2.11	18	8.74	1.73	19	9.47	2.33	32	9.53	1.74	32
Junior high	9.95	1.68	19	8.11	1.33	19	8.94	1.66	18	8.28	1.60	18
Arithmetic												
Elementary	8.63	2.43	19	7.58	1.98	19	9.34	2.86	32	8.94	2.83	32
Junior high	7.47	1.54	19	8.00	1.60	19	8.83	2.26	18	8.00	1.64	18
Change score *M*												
Comprehension												
Elementary	1.42	3.22	19	1.26	2.13	19	.53	2.72	32	.41	2.67	32
Junior high	−.17	2.28	18	1.11	2.52	18	.67	2.54	18	1.56	2.04	18
Vocabulary												
Elementary	.22	1.80	18	.84	1.83	19	1.16	2.38	32	1.31	2.52	32
Junior high	1.05	1.54	19	.37	1.54	19	.00	.91	18	.78	1.83	18
Arithmetic												
Elementary	−.11	2.60	19	2.00	1.91	19	.00	2.74	32	.59	2.80	32
Junior high	.63	1.57	19	1.32	1.80	19	−.28	1.18	18	−.39	1.46	18

disadvantaged youngsters; the special feature of these data is that both advantaged and disadvantaged samples were drawn from learning problem populations and were equated for severity of learning deficit.

Although there appear to be some sizable differences in amount of vocabulary change, the vocabulary fluctuations are very variable and none of these differences are significant. The possible cultural bias of the vocabulary scale may make it relatively insensitive to vocabulary increments in disadvantaged populations.

No such limitation applies to the arithmetic subscale of the WISC. However, there are pretest differences on this measure that could have an influence on the change scores. These differences occur largely at the elementary level, the pretest mean of the Fernald disadvantaged children being significantly lower than that of either of the other disadvantaged groups. This pretest difference is, in part, a consequence of the fact that some of the duller students left the enrichment and control samples during the course of the study, thereby elevating the mean score of the remaining children. Regardless of these initial differences, at both the junior high and elementary levels, the Fernald disadvantaged boys show a significant increase in arithmetic performance, which is reliably greater than that achieved by the other disadvantaged groups (based on both analyses of variance and covariance). Also, the increment is significantly greater than the change in the advantaged elementary school boys. The gains manifested by the Fernald disadvantaged groups on this arithmetic subscale can be viewed as an increment in IQ. The obvious connection between these changes and the increments found on the arithmetic achievement subtests points to the more general relationship between IQ and achievement and the often arbitrary distinction made between these two concepts.

To determine the differences between the advantaged and disadvantaged groups, the

scores from the six subtest short forms of the WISC were used. The means are somewhat lower than in the case of the IQ data previously reported for the entire sample, but the size of the differences between the advantaged (elementary group = 96.5; junior high = 95.4) and disadvantaged (elementary = 89.5; junior high = 89.7) groups is comparable to the data cited in Table 4 and may be explained in the same way. Turning to the subtest scores, the largest differences between the Fernald disadvantaged and advantaged boys, and the only individually reliable ones, were on the vocabulary and similarities subscales ($p < .0005$ and $p < .005$, respectively), both of which entail a high verbal factor. (While the comprehension subtest is also a matter of verbal understanding, it does not require verbal definition as is the case for vocabulary and similarities.) However, since the sample is small and the array of measures is limited, it would be premature to conclude that intellectual differences between advantaged and disadvantaged learning problem populations lie primarily in the area of verbal proficiency.

Perceptual–cognitive testing. The Wepman Auditory Discrimination Test and the Bender Gestalt measure of perceptual–motor functioning were administered at the beginning and at the end of the academic year in order to determine whether changes in conceptual, academic skills were accompanied by systematic changes at the perceptual level. These tests were given only to the elementary age children, since they are not appropriate to older age groups except where brain damage or related nervous system malfunctioning is suspected. The means of the pretest and change error scores for both tests are presented in Table 5.

TABLE 5

WEPMAN AUDITORY DISCRIMINATION TEST AND BENDER VISUAL GESTALT TEST MEANS (ELEMENTARY STUDENTS ONLY)

Test	Measure	Fernald advantaged	Fernald disadvantaged	School enrichment	Control
		Pretest *M*			
Wepman	*M*	1.89	3.49	3.83	3.54
	SD	1.49	2.74	2.00	2.60
	n	19	19	29	28
Bender	*M*	3.11	4.37	4.42	4.57
	SD	2.74	3.48	3.02	2.62
	n	18	19	31	28
		Change score *M*			
Wepman	*M*	.0	−.79	−1.34	−.61
	SD	2.16	3.12	2.39	2.73
	n	19	19	29	28
Bender	*M*	−.61	−.84	−.77	−1.07
	SD	1.69	3.56	2.62	2.73
	n	18	19	31	28

The mean error score for the advantaged children on the Wepman is significantly lower than the mean error score obtained by each of the disadvantaged groups. All of the disadvantaged groups declined in error scores; this change, however, is not significantly different from the zero mean change score of the advantaged group. The pretest difference may well reflect a vocabulary difference rather than one of auditory capacity or "tuning out," inasmuch as familiarity with the words used in the Wepman would influence the error score. The Bender Gestalt data are minimally influenced by any verbal component. The child simply has to copy a figure, verbalizations only entering into the instruction for this test. The child's productions were scored by the Koppitz method, higher scores reflecting more errors. Although none of the pretest differences are significant, the differences are in the same direction as obtained on the Wepman. (A similar result was obtained on the Frostig Developmental Test of Visual Perception.) If the three disadvantaged groups are combined and then compared to the advantaged children, the difference is statistically reliable. As in the case of the Wepman, there are no significant differences in change scores.

These data, then, indicate that the experimental program had no significant effect upon these perceptual–cognitive skills that have been linked to learning problems, particularly in reading. While one cannot conclude from these data alone that changes on the perceptual–cognitive level were irrelevant to changes in academic skills, this inference is certainly a reasonable one. The

poorer performance of the disadvantaged as compared to the advantaged children requires additional analysis and supplementation by other data before it can be adequately interpreted.

Attitudinal testing. The present brief discussion of our findings with regard to group attitudinal differences is limited to locus of perceived control and evaluation, vocational aspirations, and ethnic attitudes. The last two instruments were readministered at the end of the year to assess changes in attitudes. There is some evidence that disadvantaged populations are more likely to perceive themselves as controlled by external and accidental forces than are middle- and upper-income groups (a not necessarily inaccurate perception). More germane to the present study, the Coleman report suggests that one of the best predictors of the disadvantaged child's response to special experiences is the extent to which he feels that he can control his own fate. At the beginning of the second and third years of the project, we administered the locus of control and locus of evaluation instruments. The former scale assesses the degree to which the child feels that he, himself, versus external forces has control over his behavior, while the latter focuses on the basis for his self-evaluation. Higher scores on these scales reflect greater feelings of autonomy, self-reliance, and self-control.

We found clear evidence of an age factor, the elementary boys obtaining reliably lower internalization scores on both instruments than the junior high boys. Second, we found very little difference at the elementary level between the advantaged and disadvantaged groups, but a number of interesting differences emerge at the junior high level on the Locus of Control Scale. The advantaged group obtained significantly higher scores than did the controls ($F = 4.6, p < .05$) and also obtained higher scores than did the enrichment group, although the latter difference fell short of statistical significance. What is most interesting, however, is the finding that the Fernald disadvantaged boys obtained reliably higher scores than those obtained by either the control ($F = 6.6, p < .025$) or enrichment groups ($F = 4.3, p < .05$). This Locus of Control measure was administered only a few weeks after the initiation of the experimental program, and these differences suggest that the simple exposure of the disadvantaged boys to the Fernald School resulted in stronger feelings of autonomy and a greater acceptance of personal responsibility for one's own performance and actions.

A measure of vocational aspirations was administered as one means of determining whether the Fernald and enrichment experiences produced any changes in the child's perception of the opportunities available to him and the level of vocational goals he set for himself. The advantaged children, as might be expected, were found to have higher initial aspirations than the disadvantaged boys. However, the differences were not large, reflecting perhaps the fact that the advantaged children perceive themselves as having learning problems that limit their vocational possibilities. We also found that the junior high children have more ambitious vocational aspirations than their elementary counterparts, which is encouraging since it suggests that the junior high boys have not become overwhelmed and completely discouraged by their school failures.

The change scores revealed very few reliable differences. The vocational measure reflected a lowered level of aspiration for all of the junior high groups on retesting. The experimental program, then, was not effective in raising the aspirations of the Fernald junior high boys, despite the gains they made in academic skills. Perhaps the boys were only being more realistic on retesting. At the elementary level, the Fernald disadvantaged children showed an elevation in aspiration reliably greater than the change in the advantaged subjects, but since they were the lowest group initially, the change was not large enough to bring them in line with the other groups.

One of the questions that was of central interest to us concerned the effects of the integration experience upon the child's perceptions of his own and other ethnic groups. Because of school policies, we used a rather

indirect procedure that entailed presentation of photographs of Anglo (Caucasian), black and Mexican-American children and elicitation of judgments of characteristics of the children in the photographs.

Both groups, in their initial judgments, gave higher rankings on kindness to the Anglo photographs, particularly at the junior high level. The Anglo photographs were also ranked much higher by both advantaged and disadvantaged groups for "happiest boy" and boy who gets "best grades." On the dimension of best grades, the junior high age group gave considerably higher rankings to the Anglo stimuli than did the elementary boys, whether advantaged or disadvantaged.

It is only on the traits denoting physical skills that the Anglo stimulus boys were given lower ranks than the black and Mexican-American stimuli. The Mexican-Americans were judged as the strongest especially by the advantaged boys, with the black stimuli falling close behind. The judgments of "fastest boy" were much the same, with blacks and Mexican-Americans receiving similar ranks and the Anglo boys seen as less fast than the others, especially by the elementary groups.

What is particularly striking about these data is the extent to which the Anglo advantaged children and the largely black disadvantaged children shared a common conception of the relative attributes of Anglos, blacks, and Mexican-Americans. In terms of the child's overall self-image, it would be interesting to know the relative importance of these traits for the advantaged and disadvantaged child.

The ethnic attitude change scores were not very illuminating and, in some respects, were rather disappointing. There were very few significant differences between the Fernald disadvantaged group and the enrichment and control groups in the degree and direction of change shown. In their rankings of "kindest boy," the Fernald disadvantaged elementary group increased their ranking of the black stimuli while at the same time lowering the rankings of the Anglo photographs. The corresponding changes in the elementary enrichment and control groups were directly opposite in direction. Again, at the elementary level, both Fernald groups saw the black child as happier on retesting than did the enrichment and control groups. However, the differences were reliable only for the Fernald advantaged group comparisons. Also at the elementary level, the Fernald disadvantaged children lowered their rankings of the Anglo stimuli in judging fastest boy, while elevating the rankings of the Mexican-American and black stimuli in compensating for this shift. There were no significant differences among the various groups in the changes observed in their rankings of boy with best grades and "strongest boy."

These data provide some evidence of a positive change in the Fernald disadvantaged elementary children in the way in which they view members of their own ethnic group. However, this measure failed to reflect any reliable changes in the Fernald disadvantaged junior high level boys as compared to the other two disadvantaged groups. This is the only instance in which the disadvantaged junior high boys attending the Fernald School displayed weaker experimental effects than their elementary counterparts.

Summary and Conclusions

There are several clear-cut findings that emerge from the data, other findings that form a trend consistent with the principal results, and still other data that are only suggestive or ambiguous. The major experimental finding is clearly the increase in achievement observed in the disadvantaged children attending the Fernald School and the failure of the enrichment program to exert an influence significantly greater than that provided by the control experience. The California Achievement Test findings are buttressed by the qualitative performance of the Fernald disadvantaged group and by a significant increment on the arithmetic subtest of the WISC. These effects are generally stronger for the junior high group than for the elementary group.

One of the cognitive areas that did not reflect any experimental effects was vocabulary. Possible reasons for this have already

been discussed. In addition, the improvements noted in achievement test performance were not accompanied by significant changes in perceptual skills. The disadvantaged elementary group did show initial deficiencies in the latter area and did improve in performance, but those changes appeared to be unrelated to the experimentally produced changes in the more complex basic school skills.

The analyses of attitudinal changes yielded sporadic findings which, when significant, were consistent with the achievement test findings. On the whole, however, from these data it would appear that profound or systematic changes in these affective areas did not take place as a result of either the Fernald or the enrichment experiences.

When the disadvantaged youngster is compared with the advantaged learning disorder population, a number of cognitive differences emerge—in vocabulary and, for the younger age group, on the perceptual tasks. At the same time, there are striking areas of similarity between the two groups on other measures.

In summary, the quantitative findings and the qualitative observations suggest that black children, both the young adolescent and the elementary-school-age child from economically disadvantaged areas and with marked deficits in school achievement, can derive significant benefits from *major* variations in school inputs, for example, major reorganization of classroom, school structure, and atmosphere as contrasted with the kind of piecemeal intervention so characteristic of Title I programs. These results offer one small piece of evidence to combat the increasing pessimism concerning the efficacy of compensatory educational programs for economically disadvantaged blacks and other minorities, a pessimism that we suggest is based upon an indiscriminate analysis of Title I and related programs, disregarding the variables that distinguish effective and ineffective programs. At the same time, the data indicate that labels such as "culturally disadvantaged" and "learning disordered" do not encompass discrete diagnostic behavioral categories and can obscure important functional similarities between children assigned these diverse labels. The learning problems of an economically advantaged child require for their resolution substantial remedial efforts over a period usually measured in years. Should the learning problems presented by the economically disadvantaged child require any less?[5]

REFERENCES

Baratz, S. S., & Baratz, J. C. Early childhood intervention: The social base of institutional racism. *Harvard Educational Review*, 1970, **40**, 29–50.

Bender, L. *Visual Motor Gestalt Test.* New York: American Orthopsychiatric Association, 1946.

Carleton, F. O., & Stacey, C. L. Evaluation of selected short forms of the Wechsler Intelligence Scale for Children (WISC). *Journal of Clinical Psychology*, 1954, **10**, 258–261.

Clark, K. B. *Dark ghetto.* New York: Harper & Row, 1965.

Deutsch, C. P. Environment and perception. In M. Deutsch, I. Katz, & A. Jensen (Eds.), *Social class, race, and psychological development.* New York: Holt, Rinehart & Winston, 1968.

Deutsch, M. Nursery education: The influence of social programming on early development. *Journal of Nursery Education*, 1963, **18**, 191–197.

Dixon, W. J. (Ed.) *BMD: Biomedical computer programs, X-series supplement.* (University of California Publications in Automatic Computation 3) Berkeley: University of California Press, 1969.

Enburg, R., Rowley, V., & Stone, B. Short forms of the WISC for use with emotionally disturbed children. *Journal of Clinical Psychology*, 1961, **17**, 280–284.

Fantini, M. D., & Weinstein, G. *The disadvantaged: Challenge to education.* New York: Harper & Row, 1968.

Fernald School. *A training, demonstration, and research program for the remediation of learning disorders in culturally disadvantaged youth.* (Project report submitted in connection with Project M7-200) Sacramento: California State Department of Education, 1969.

Feshbach, S., & Adelman, H. S. An experimental

[5] During the course of our three-year study, four of the disadvantaged black children were kept a second year at the Fernald School. Of the two elementary-age children, one gained .4 years in total grade placement score the first year and 2.0 years the second year, while the corresponding scores for the other child were .5 years and 1.2 years. For the two junior high age boys, the gains in grade placement for the first and second years were 1.1, 1.2, and 1.4, 1.5, respectively.

program of personalized classroom instruction in disadvantaged area schools. *Psychology in the Schools,* 1971, **8,** 114–120.

FROSTIG, M., LEFEVER, D. W., & WHITTLESEY, J. R. B. *Administration and scoring manual for the Marianne Frostig Developmental Test of Visual Perception.* Palo Alto, Calif.: Consulting Psychologists Press, 1966.

GERARD, H. B. *Factors contributing to adjustment and achievement in racially desegregated public schools.* (Progress report for the National Institute of Mental Health) Rockville, Md.: National Institute of Mental Health, 1969.

HESS, R., SHIPMAN, V., BROPHY, J., & BEAR, R. *The cognitive environments of urban preschool children.* Chicago: University of Chicago, 1968.

HUNT, J. MCV. Environment, development, and scholastic achievement. In M. Deutsch, I. Katz, & A. Jensen (Eds.), *Social class, race, and psychological development.* New York: Holt, Rinehart & Winston, 1968.

JAMES, W. Internal versus external control of reinforcements as a basic variable in learning theory. Unpublished doctoral dissertation, Ohio State University, 1957.

JENSEN, A. How much can we boost IQ and scholastic achievement? *Harvard Educational Review,* 1969, **39,** 1–123.

KEMPTHORNE, O., ET AL. *Analysis of variance procedures.* Wright-Patterson Air Force Base, Ohio: Aeronautical Research Laboratory, 1961.

ORNSTEIN, A. C., DOLL, R. C., ARNEZ, N. L., & HAWKINS, M. *Educating the disadvantaged: School year 1969–1970.* New York: AMS Press, 1971.

SCHEFFÉ, H. *The analysis of variance.* New York: Wiley, 1959.

TIEGS, E. W., & CLARK, W. W. *California Achievement Test.* Monterey: California Test Bureau, 1957.

WEPMAN, J. M. *Auditory Discrimination Test.* Chicago: Author, 1958.

WRIGHTSTONE, J. W., FORLANO, G., FRANKEL, E., LEWIS, B., TURNER, R., & Bolger, P. *Evaluation of higher horizons program for underprivileged children.* (Cooperative Research Project No. 1124) New York: Board of Education of the City of New York, Bureau of Educational Research, 1964.

(Received November 27, 1972)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 29-32

ATTENTION AND READING ACHIEVEMENT IN FIRST-GRADE BOYS AND GIRLS[1]

S. JAY SAMUELS[2] AND JAMES E. TURNURE

University of Minnesota

A behavior observation schedule was utilized to investigate sex differences in classroom attentiveness and the relationship of such attentiveness to reading achievement among first-grade children, 74 boys and 58 girls. Girls were found to be significantly ($p < .01$) more attentive than boys and to achieve higher word recognition scores ($p < .05$). Further, word recognition was found to be significantly ($p < .01$) related to attentiveness for the group as a whole, with reading readiness controlled in a covariance analysis. This latter finding replicates previous results with fourth- and sixth-grade pupils but demonstrates that the relationship obtains with beginning readers before a history of academic success–failure has been established.

The purpose of this study was to determine whether there are sex differences in classroom attentiveness and to determine if attention (i.e., visual orienting behavior or direction of gaze [Turnure, 1970, 1971]) is related to reading achievement.

Numerous studies in reading have reported that in naturalistic classroom settings, reading achievement of girls is superior to that of boys (Dykstra, 1967; Gates, 1961). However, in the laboratory, where attentional behavior of the subject is more easily controlled, sex differences in reading-analogous paired-associate learning are not found (Jeffrey & Samuels, 1967; Peterson, 1972). There have been some demonstrations that the performance of boys is superior to that of girls on a reading-type task under certain classroom conditions (McNeil, 1964); in the McNeil study, orienting behavior was controlled through placing students in cubicles and utilizing headphones for audio input.

When asked why students succeed in a particular subject matter, teachers tend to attribute success to teacher-controlled variables such as methodology. On the other hand, academic failure is generally explained by reference to intrastudent variables, such as lack of intelligence, readiness, motivation, or attention (Baldwin, Johnson, & Wiley, 1970).

A recent study by Lahaderne (1968) reported that school achievement at Grade 6 was related to attention. A similar finding was reported by Cobb (1972) for fourth-grade pupils. However, when measurements come this late in the child's academic career, one cannot be sure if inattentiveness may not represent avoidance behavior resulting from academic failure. That is, lack of school success may lead to inattentiveness rather than the reverse, since lack of reinforcement in school settings generally leads to extinction of task-relevant behaviors.

By replicating Lahaderne's study in Grade 1, we hoped to determine whether attentiveness was related to academic achievement (i.e., reading) prior to the effects of long-term success–failure school experiences. In addition, the study was designed to determine if the expected superior

[1] The authors would like to express their appreciation for the cooperation extended them by the pupils and teachers of the Minneapolis public school system. We also acknowledge the contributions of Karen Anderson, Bridgitte Schroeder, and Jenna Beard as observers and of Martha Thurlow, Sharon Larsen, and Teara Archwamety, who assisted with data processing analysis.

[2] Requests for reprints should be sent to Jay Samuels, Minnesota Reading Research Project, Center for Research in Human Learning, University of Minnesota—Elliott Hall, Minneapolis, Minnesota 55455.

reading achievement of girls was related to observed attentiveness in the classroom.

Method

Subjects

Eighty-eight first graders, 53 boys and 35 girls, were observed. The subjects were obtained from four classrooms in two middle-class schools in the Minneapolis school system. The teachers in these schools used traditional three-group reading methods and basal reader materials.

Procedure

An observer was assigned to each of the four classrooms to record the attentional behaviors of the pupils during the reading hour. Attention was defined and measured in a manner similar to that reported by Lahaderne (1968). Task-relevant behaviors (i.e., orienting eyes to text or teacher, working on reading follow-up exercises, observing chalkboard or overhead projection, or otherwise following the instructional directions of the teacher) were scored as positive instances of attentiveness. Negative attentiveness consisted of nontask-orienting behavior such as failing to follow instructional directions, closing eyes, working or playing with nonassigned materials, etc.

Attentive and inattentive behaviors were recorded on a scoring sheet that listed the children's names according to their reading groups. Each child was observed in sequence in accordance with his listing on the scoring sheet. A six-second scoring method was used. A child was observed for four seconds, and in the next two seconds a plus (+) or minus (−) was entered on the scoring sheet representing the observer's judgment of task attentiveness or inattentiveness, respectively. Question marks signified ambiguous instances in which it was unclear to the observer whether or not the pupil was attentive. It was possible to record 600 observations during each reading hour. Observers made 15 visits to each classroom over the course of a month. The attentional data used for analysis was a proportion score comprised of the number of positive instances divided by the total number of both positive and negative instances; question marks were excluded.

Reliability

A videotape was made of first-grade children grouped around a table doing reading follow-up work. A 15-minute segment of the film was selected for training purposes, and a 2-minute segment not previously exposed was retained for the test of interrater reliability.

Using the six-second scoring method described earlier, observer reliability was calculated by dividing the total number of agreements by the total number of recorded behaviors. This resulted in a reliability of 89%, which is similar to those reported by Cobb (83%) Lahaderne (83%–100%).

Reading Achievement Measure

Reading achievement was measured by presenting 45 words, randomly selected from the Dolch (1956) list of basic sight words, for recognition. Each of these words was typed onto a 3 × 5 inch card with a primary typewriter. The words were presented individually and the student was given up to six seconds to respond. The experimenter recorded all correct responses on a score sheet. No feedback was given on the test.

Data Analysis

Reading readiness scores were available and were used as a covariate in the analysis of word recognition. These readiness scores were also used in analyzing the data by means of partial correlations. Attention scores were placed in four quartiles: Q_1 = .68 and below; Q_2 = .69–.80; Q_3 = .81–.87; and Q_4 = .88 and above. A subject, for example, was placed in quartile Q_1 if he was attentive 68% of the time or less. Attention and sex served as the two independent variables in the analysis of covariance of word recognition scores.

Results

Sex Differences in Reading Readiness and Attention

The mean reading readiness score for the boys was 67.43 (SD = 18.14), and the mean for the girls was 64.34 (SD = 27.18). Comparing these two means by t tests indicated that the difference was not significant ($t < 1$, $df = 86$).

Comparing the two sexes on attention indicated that the mean attention score for the boys was .76 (SD = .13) and the mean for the girls was .84 (SD = .10). This difference in attention was significant in favor of girls ($t = 3.08$, $df = 86$, $p < .01$).

Attention and Reading Achievement

Table 1 shows the mean word recognition scores for each of the attention categories and for the two sexes. Inspection of the table indicates that as attention increases, there is a corresponding increase in word recognition. Also, females have a higher recognition score in comparison to males. A Sex × Attention analysis of covariance of these scores, using reading readiness as the covariate, found the following: a significant main effect for attention ($F = 8.46$, $df = 3/79$, p

TABLE 1
MEAN WORD RECOGNITION SCORES AND STANDARD DEVIATIONS FOR ATTENTION CATEGORIES AND SEX

Measure	Attention category				Sex	
	Q_1	Q_2	Q_3	Q_4	Male	Female
$\bar{X}$	16.18	25.05	26.83	33.87	22.68	30.03
SD	13.04	10.24	13.38	13.49	13.74	13.68

Note. The four quartiles used are as follows: Q_1 = .68 and below; Q_2 = .69–.80; Q_3 = .81–.87; and Q_4 = .88 and above.

< .001) and a significant main effect for sex ($F = 3.96$, $df = 1/79$, $p < .05$). The Sex × Attention interaction was not significant ($F < 1$).

Newman-Keuls tests were computed to determine which of the word recognition scores were significantly different from each other for each of the four attention categories. The following was found: All comparisons were significantly different from each other ($p < .05$) with the exception of Q_2 and Q_3.

In order to facilitate comparability from the present study to the reports of Lahaderne (1968) and Cobb (1972), correlations between attention and word recognition were computed. A Pearson product-moment correlation of .44 was computed ($p < .01$). This correlation is close to the value found between attention and reading achievement by Cobb ($r = .45$) for the fourth grade and by Lahaderne ($r = .51$–$.39$) for sixth graders. The partial correlation in this study, conducted between the variables of attention and word recognition and controlling for reading readiness, was .44.

DISCUSSION

In this study, it was found that girls were significantly superior to boys in word recognition, as had been previously reported, and significantly superior in classroom attentiveness as well. We also found that increasing degrees of attention were related to superior word recognition. Thus, like Lahaderne, and more recently Cobb and Hops (1973), we found that overt task-relevant orienting behavior was related to scholastic achievement; furthermore, this relationship was obtained in beginning reading before a long history of academic failure had been established.

Environmental factors have an important effect on children's scholastic achievement. In Germany, where male teachers predominate, boys were found to be superior in reading achievement (Preston, 1962). McNeil (1964) reported that in a reading-type task, boys learned more than did girls when the learning situation was designed to approximate a mechanical manipulative activity, but there was a reversal in achievement when these children were taught reading in the classroom by women. Sex differences in verbal learning are not found in laboratory studies in which attention is carefully controlled.

One educational implication of the present study, as well as the studies reported by Lahaderne (1968) and Cobb (1972), is that contrary to teachers' opinions that only low achievement is related to intrastudent variables (Baldwin et al., 1970), both high and low reading achievement appear to be related to attention. Another implication would be that the sex difference favoring girls frequently found in reading achievement seems to be mediated by an attentional variable. Behavior modification literature indicates that orienting behavior can be controlled by the teacher (Packard, 1970; Walker & Buckley, 1968) and that such control of attention can be conducive to higher academic achievement (Cobb & Hops, 1973). Since the time of James (1890), psychologists have believed that attention is the sine qua non for learning. It should be pointed out to teachers that reading achievement has been found to be related to attention. Instructional success clearly requires that teachers secure and maintain the attention of all their pupils. In light of the finding that teachers tend to attribute academic failure to intrastudent variables, the viewpoint expressed by Goldiamond and Dyrud seems appropriate:

> The performance of the student may be to a considerable extent a function of the procedures used to establish that behavior; we should look to deficits in our own procedures before ascribing deficits to the students or difficulty to the problem [1966, p. 99].

REFERENCES

BALDWIN, T. L., JOHNSON, T. J., & WILEY, D. E. The teacher's perception and attribution of causation. Paper presented at the meeting of the American Educational Research Association, Minneapolis, April 1970.

COBB, J. A. Relationship of discrete classroom behaviors to fourth-grade academic achievement. *Journal of Educational Psychology*, 1972, **63**, 74–80.

COBB, J. A., & HOPS, H. Effects of academic survival skill training on low achieving first graders. *Journal of Educational Research*, 1973, **67**, 108–113.

DOLCH, E. W. School research in reading. *Elementary English*, 1956, **33**, 76–80.

DYKSTRA, R. *Continuation of the coordinating center for first-grade reading instruction programs.* (Project 6-1651, Contract OEC 3-7-001651-0472) Washington, D. C.: United States Office of Education, 1967.

GATES, A. I. Sex differences in reading ability. *The Elementary School Journal*, 1961, **61**, 431–434.

GOLDIAMOND, I., & DYRUD, J. E. Reading as operant behavior. In J. Money (Ed.), *The disabled reader: Education of the dyslexic child.* Baltimore, Md.: Johns Hopkins Press, 1966.

JAMES, W. *Principles of psychology.* New York: Holt, 1890.

JEFFREY, W. E., & SAMUELS, S. J. Effect of method of reading training on initial learning and transfer. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 354–358.

LAHADERNE, H. M. Attitudinal and intellectual correlates of attention: A study of four sixth-grade classrooms. *Journal of Educational Psychology*, 1968, **59**, 320–324.

MCNEIL, J. D. Programmed instruction versus usual classroom procedures in teaching boys to read. *American Educational Research Journal*, 1964, **1**, 113–119.

PACKARD, R. G. The control of "classroom attention": A group contingency for complex behavior. *Journal of Applied Behavior Analysis*, 1970, **3**, 13–28.

PETERSON, J. Effects of sex of E and sex of S in first and fifth grade children's paired-associate learning. *Journal of Educational Research*, 1972, **66**, 81–84.

PRESTON, R. Reading achievement of German and American children. *School and Society*, 1962, **90**, 350–354.

TURNURE, J. E. Children's reactions to distractors in a learning situation. *Developmental Psychology*, 1970, **2**, 115–122.

TURNURE, J. E. Control of orienting behavior in children under five years of age. *Developmental Psychology*, 1971, **4**, 16–24.

WALKER, H. M., & BUCKLEY, N. K. The use of positive reinforcement in conditioning attending behavior. *Journal of Applied Behavior Analysis*, 1968, **1**, 245–250.

(Received February 16, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 33-39

RELATIONSHIPS OF STATE AND TRAIT ANXIETY TO CONCEPT-LEARNING PERFORMANCE

JOEL MEYERS[1] AND ROY MARTIN

Temple University

This experiment investigated relationships between anxiety and concept learning. Two hypotheses were examined: (*a*) A-State would have a stronger debilitating effect on concept-learning performance than A-Trait and (*b*) task conditions would affect A-State. Sixty-one undergraduate subjects were randomly assigned to either high- or low-ego-involving conditions. All subjects performed extradimensional-shift concept-learning tasks. The performance of high-A-State subjects was significantly inferior to that of low-A-State subjects, while there were no differences in performance between high- and low-A-Trait subjects. Furthermore, high-ego-involving instructions increased A-State and task completion reduced A-State for these subjects. These data support Spielberger's state–trait interpretation of anxiety.

Research investigating the relationship between anxiety and concept learning has produced inconsistent results and it has been proposed that factors such as task complexity, ego-involving conditions, and cognitive abilities be examined in an effort to clarify these results (Denny, 1966; Meyers & Dunham, 1972). Differentiating between trait and state anxiety might also help to produce more consistent research in this area, since there is extensive evidence that measures of anxiety that are more specific to the situation relate more strongly to learning performance than general measures. For example, Sarason's research (Sarason & Palola, 1960) suggests that test anxiety has a stronger relationship to performance on testlike tasks than trait anxiety. Similarly, Spielberger's (1966) research indicates that state anxiety (feelings of apprehension and heightened autonomic nervous system activity) is more likely to relate significantly to learning performance than trait anxiety (a generalized predisposition to experience state anxiety).

The drive theory interpretation of anxiety (Spence & Spence, 1966) has been employed to explain the relationships between anxiety and concept-learning performance (Denny, 1966; Maltzman, Fox, & Morrissett, 1953). According to this position, the learning performance of high-anxious subjects will be inferior to that of low-anxious subjects on tasks with competing responses (i.e., incorrect response tendencies that are higher in the habit hierarchy than correct response tendencies). On tasks low in competing responses (i.e., where the correct response is dominant in the habit hierarchy relative to incorrect responses), it is predicted that performance of high-anxious subjects will be superior to that of low-anxious subjects. However, it is not easy to assess the effects of anxiety in terms of drive theory, because it is difficult to determine accurately the number of competing responses.

Perhaps drive theory could be more adequately evaluated by using extradimensional-shift concept-learning tasks, since it may be possible to estimate the number of competing responses associated with these tasks. This type of problem involves an unannounced shift in the solution, as the responses that are originally learned are incorrect in the final portion of the task. Consequently, the incorrect response tendencies should be higher in the habit hierarchy than the correct responses.

[1] Requests for reprints should be sent to Joel Meyers, Department of School Psychology, Temple University, College of Education, Philadelphia, Pennsylvania 19122.

In one previous investigation using this type of task, Meyers and Dunham (1972) found no significant main effects for anxiety on concept-learning performance. However, these researchers measured trait rather than state anxiety. For the reason outlined above, it is hypothesized that state anxiety would be a more adequate measure of anxiety as a drive. Consequently, the present investigation is an extension of the one reported by Meyers and Dunham in that state anxiety was assessed in addition to trait anxiety.

Whereas much important work investigating the relationship between anxiety and learning has been oriented to determining the effects of anxiety on learning performance, recent investigations have also begun to determine effects of the task on anxiety experienced by subjects. For example, Martin (1970) reported that qualifying examinations for doctoral candidates created substantial levels of state anxiety and that the degree of anxiety increased as the date of the exam approached. In addition, Martin and Meyers (1972) have shown that the level of state anxiety increased as the date of a final examination became more proximate. Finally, O'Neill, Spielberger, and Hansen (1969) found that the level of difficulty of a task had a significant effect on the state anxiety experienced by subjects.

In view of the literature cited regarding the effects of state and trait anxiety on concept-learning performance and the literature regarding the effects of task conditions on state anxiety, the purpose of this investigation was twofold: (*a*) to determine whether state anxiety would have a stronger debilitating effect on concept-learning performance than trait anxiety and (*b*) to determine whether certain aspects of the experimental situation would have a significant effect on the level of state anxiety experienced by subjects.

Method

Subjects

Subjects were 61 introductory educational psychology students at the University of Texas. They were randomly assigned to two experimental treatments, which consisted of either high- or low-ego-involving instructions (adapted from Sarason, 1956). There were 32 subjects in the high-involvement condition and 29 subjects in the low-involvement condition. Originally, there were 66 subjects; however, 5 subjects who did not complete the task were not included in the analysis.

Problem

Unidimensional concept-learning tasks were used in this experiment. There were two samples followed by two experimental problems, and these tasks were graded in difficulty. The first sample had only one dimension and it was relevant to the solution; the second sample contained three dimensions with one dimension relevant to solution; and the experimental problems consisted of five dimensions with one relevant to solution of the task.

The experimental-concept task was an extra-dimensional-shift task involving negative transfer, which consisted of two consecutive unidimensional, four-category problems with five dimensions. The first of these two problems is referred to as the original-learning problem, and the second problem is referred to as the transfer-learning problem.

A deck of 1,024 3 × 3 inch cards was derived from all possible combinations of the words found in Table 1 with the restriction that every card contain one word from each of five dimensions. These five dimensions included tools, fruit, vehicles, clothes, or furniture. There were four possible words representing each dimension, and these are the words found in Table 1. The placement of dimensions on each card was random so that no dimension would become associated with a particular position. The cards were placed in a random order, and to insure a random sequence of stimuli, starting points were chosen randomly for each subject.

While looking at a series of the above stimulus cards, one at a time, the subject's task was to determine into which of four possible response categories (1, 2, 3, or 4) each observed card belonged. In reference to Table 1, either tools, fruit, vehicles, or clothes was the dimension relevant to solution on the original-learning problem. For example, if tools was the relevant dimension, the subject may have had to learn that hammer = 1, pliers = 2, nails = 3, and saw = 4, and in this instance the

TABLE 1

Stimulus Materials in the Concept-Learning Problem

Dimension				
Tools	Fruit	Vehicles	Clothes	Furniture
hammer	cherry	plane	hat	bed
pliers	banana	train	shoes	couch
nails	apple	car	shirt	table
saw	pear	boat	pants	chair

remaining four dimensions (fruit, vehicles, clothes, and furniture) would be irrelevant to the solution of the problem.

The same deck of cards was used in the transfer-learning problem. While either tools, fruit, vehicles, or clothes was the dimension relevant to solution of the original-learning problem, the transfer problem involved a shift to a different dimension (furniture) as the basis to solution of the problem. Therefore, on the transfer-learning problem, subjects had to learn different response associations such as sofa = 1, table = 2, desk = 3, and chair = 4.

Procedure

All of the data were collected during individual testing sessions (i.e., one session per subject). During each session the subject was asked to fill out a short form (State Anxiety 1) of the Spielberger State Anxiety Inventory followed by the Spielberger Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970). The short form of the state anxiety inventory was used so that when state anxiety was measured during the task, the measurement would produce minimal interference with performance.

After completing the trait anxiety inventory, the experimenter explained the nature of the task to the subject. Then the two sample problems were presented. Each sample and the experimental problem were administered by placing one stimulus card at a time on the table in view of the subject. After the subject indicated which of the four response categories (1, 2, 3, or 4) he thought was appropriate for each card, the experimenter indicated whether the response was correct or incorrect. When the response was wrong, the experimenter stated the correct answer.

Following a criterion of eight consecutive correct responses on each of the two samples, all subjects received one of the following sets of instructions designed to produce either high or low involvement with the task.

High ego-involvement:

> Do I have your name spelled correctly? How old are you? What is your major? What classes are you in? What is your grade point average? Do you expect to get a Bachelor's degree? The first problem was a practice problem. Now we will begin the next problem. This is similar to a short-form intelligence test in that it involves the ability to think in abstract terms. Pay close attention to each card since each one missed lowers your score when it is compared with other people your age.

Low ego-involvement:

> Your performance was fine on this problem. Now I am going to give you another practice problem. What concerns me is not how well you do, but rather the characteristics of the problem which are uncovered. This is a common procedure in this sort of experiment.

The second measure of state anxiety (A-State 2) was administered immediately following these instructions, and then the experimental concept-learning task was begun by presenting the original-learning problem. When a criterion of 13 consecutive responses was reached, there was an unannounced shift in the solution and the transfer-learning problem was begun. Rather than one of the other four dimensions, furniture was now the dimension relevant to solution. Performance was terminated when a criterion of 13 consecutive responses was reached on the transfer problem. State anxiety was assessed at three additional points during the task, producing a total of five A-State measures: early during the original-learning problem (A-State 3), early during the transfer problem (A-State 4), and late in the performance of the transfer problem (A-State 5).

Results

Effects of Anxiety on Performance

There were two treatment groups derived from the two ego-involving conditions, and multiple linear regression procedures were used to determine the relationship between each anxiety variable and performance in these two conditions. The number of errors on the original- and on the transfer-learning problems were the two measures used to assess performance.

The regression analyses included two steps. First, an F test was computed to determine whether the relationship between anxiety and performance was significantly different in the two treatment conditions. Since no significant differences were found between treatments, the subjects from these two conditions were combined into one group and a second analysis was completed to determine whether there was a significant relationship between anxiety and performance. This relationship was defined in terms of the slope of the regression line of anxiety on performance, and an F test was used to determine whether the slope was significantly different from zero.

Original learning. The first three measures of state anxiety were administered prior to or during the early portions of the original-learning problem, and consequently it was predicted that these three state anxiety measures would have significant debilitating

TABLE 2
MULTIPLE LINEAR REGRESSION ANALYSIS

Anxiety measure	Original-learning problem		Transfer-learning problem	
	F^a	r^2	F^a	r^2
Trait anxiety	.10	.002	1.82	.030
State 1	8.09***	.121	3.17*	.051
State 2	4.74**	.074	2.90*	.047
State 3	4.15**	.066	2.55	.041
State 4	1.95	.032	1.90	.031

[a] $df = 1/59$.
* $p < .10$.
** $p < .05$.
*** $p < .01$.

effects on performance of the original-learning problem. Table 2 reveals that this prediction was confirmed, as A-States 1, 2, and 3 were all significantly related to performance of the original-learning problem. In contrast to state anxiety, it was predicted that trait anxiety would not have a significant relationship to performance, and Table 2 reveals that this expectation was also confirmed.

The relationships between anxiety and performance of the original-learning problem are depicted graphically in Figure 1 in order to help describe these results more clearly. For example, it is clear that there is no relationship between A-Trait and performance, as the number of errors is essentially the same for subjects one standard deviation above the mean (high A-Trait) and for subjects one standard deviation below the mean (low A-Trait) in A-Trait.

In contrast to trait anxiety, Figure 1 reveals that the significant relationships between the first three measures of state anxiety and original-learning performance were all clearly debilitating. Subjects scoring one standard deviation above the mean in state anxiety (high A-State) had significantly more errors on the original-learning problem than subjects scoring one standard deviation below the mean in state anxiety (low A-State). Consistent with predictions, Figure 1 demonstrates that there were debilitating relationships between A-State and original-learning performance that were not found for A-Trait in this study.

Finally, the fourth measure of state anxiety was administered during the early portions of the transfer-learning problem, after the original-learning problem had been completed. Consequently, a significant debilitating relationship between this variable and performance of the original-learning problem was not expected. Table 2 revealed that this expectation was also confirmed.

Transfer learning. The comparisons between state anxiety and trait anxiety are less clear during the transfer-learning problem. Similar to original learning, Table 2 reveals that trait anxiety did not relate significantly to performance of the transfer problem. However, Table 2 also reveals that there were no significant relationships to transfer performance for any of the measures of A-State, although the relationships for A-States 1 and 2 did approach significance. While not significant, the direction of these relationships was such that they indicated debilitating effects for A-State and A-Trait on transfer learning.

Factors Affecting State Anxiety

In addition to investigating the effects of anxiety on concept-learning performance, an equally important goal was to examine

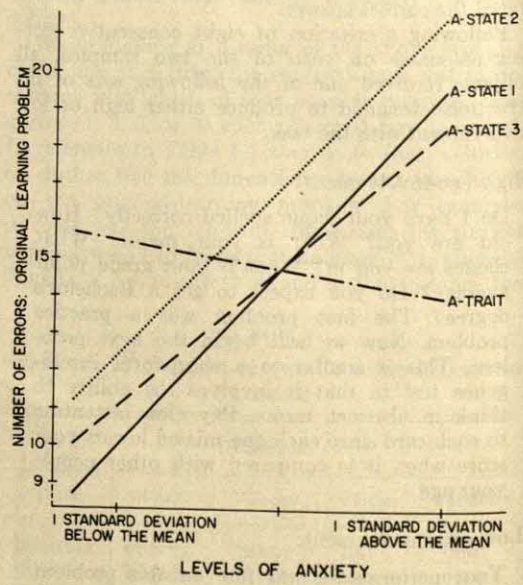


FIG. 1. Relationship between each of four anxiety measures and number of errors on original learning.

the effects of the experimental situation on the state anxiety experienced by subjects. Specifically, it was felt that task performance, task intervals (i.e., the point in time during the task when A-State is assessed), and the ego-involving instructions would have some effects on state anxiety.

The relationships between task performance and state anxiety were examined with a Pearson product-moment correlation between performance of the transfer-learning problem and subjects' scores on the fifth measure of A-State. Since this fifth measure was taken when the task was almost finished, it was assumed that performance of the transfer task had an effect on the level of anxiety rather than the reverse. This was the rationale for using a correlational analysis rather than the multiple regression procedures used with the preceding four measures of A-State. As anticipated there was a moderate relationship ($r = .35, p < .01$), which was negative in the sense that poor performance was associated with high anxiety and that performance improved as anxiety decreased. However, it was found that this relationship was dependent on the ego-involving instructions.

Specifically, for subjects with high-ego-involving instructions the negative relationship between performance and state anxiety was predictably high ($r = .49, p < .01$). On the other hand, an unexpected finding was the lack of relationship between performance and A-State for subjects in the low-ego-involving condition ($r = .14, ns$).

A repeated-measures analysis of variance was computed in order to further delineate the effects of the experimental situation on A-State. More specifically, this analysis was designed to consider the effects of both ego-involving instructions and task intervals on A-State. Consequently, the two groups of subjects receiving high- or low-ego-involving instructions formed the two treatment conditions, and the five administrations of the A-State questionnaires accounted for trials, as these anxiety scores were the dependent measures.

There was a significant main effect for trials on the state anxiety measure ($F = 5.04, df = 4/236, p < .001$), and this indicated that A-State increased following the ego-involving instructions, maintained a high level throughout task performance, and decreased as completion of the task approached.

There was also a significant Ego Involvement × Trials interaction ($F = 3.76, df = 4/236, p < .01$), and this finding provides an important elaboration to the main effect. Consistent with the above results, an inspection of the means in Table 3 reveals that under high-ego-involving instructions, state anxiety increased subsequent to the instructions, maintained this high level throughout the task, and decreased as task completion approached. On the other hand, for subjects receiving low-ego-involving instructions, Table 3 reveals that there was essentially no change in A-State. Specifically, for these subjects there was no increase in A-State following the instructions, and there was no decrease in A-State as completion of the task approached. Consequently, both the correlational data and the repeated-measures analysis of variance support the notion that task conditions have an impact on state anxiety and that ego-involving instructions can modify the relationships between the task and state anxiety.

Discussion

One important goal of this investigation was to continue recent attempts to broaden anxiety research by considering task performance and experimental conditions as they affect state anxiety (Martin, 1970; Martin & Meyers, 1972; O'Neill et al., 1969), rather than limiting questions to the

TABLE 3

Group Means and Standard Deviations for State Anxiety

Involvement	State				
	1	2	3	4	5
High					
M	7.8	9.7	9.7	9.5	8.8
SD	2.5	3.5	3.1	3.5	3.1
Low					
M	8.8	8.8	8.8	8.9	8.7
SD	2.9	3.1	2.9	3.3	3.1

effects of anxiety on learning performance. This newer area of research is becoming more feasible as instruments such as Spielberger's State Anxiety Inventory continue to be developed, and the available research, including the present study, indicates that this is a productive direction for further investigation.

Specifically, in this study there were complicated relationships between the experimental situation and A-State. It was expected that the ego-involving instructions would have an impact on A-State, and this occurred in two ways. First, as expected, high-ego-involving instructions increased A-State; whereas low-ego-involving instructions did not have this effect. Second, low-ego-involving instructions consistently eliminated the effects of task on state anxiety. For example, as was predicted from past research (Martin, 1970; Martin & Meyers, 1972), there was a negative relationship between task performance and A-State. However, this relationship occurred for subjects with high-ego-involving instructions and not for those with low-ego-involving instructions. Similarly, for subjects with high-ego-involving instructions there was a decrease in anxiety as completion of the task approached, and this confirms prior results (O'Neill et al., 1969). However, this effect was not obtained for subjects receiving low-ego-involving instructions in the present study.

It is potentially important that the effects of the task on A-State were obtained under relatively threatening conditions and eliminated when less threatening conditions were employed. An implication of this finding is that there may be ways to control the anxiety-provoking aspects of testlike situations by altering the instructions given for the test. Research would, of course, be necessary to confirm and elaborate this finding.

The second goal of the present investigation was to determine whether state anxiety would be more strongly related to performance of a concept-learning task than trait anxiety. The significant relationships between state anxiety and performance of the original-learning problem combined with a lack of significant relationships between trait anxiety and performance confirmed this expectation and provided strong support for Spielberger's state–trait interpretation of anxiety.

The results for state anxiety were in the predicted direction; that is, the performance of high-anxious subjects was inferior to that of low-anxious subjects, and this supports the Spence interpretation of anxiety as a drive (Spence & Spence, 1966). However, the transfer problem was designed to provide the strongest test of drive theory, since the unannounced shift in solution of the task was expected to create high numbers of competing responses. State anxiety did not significantly relate to transfer performance, so the predictions of drive theory were not fully supported.

While the results obtained during transfer learning are not supportive of drive theory, additional work would be needed to assess drive theory in this context. In order to produce a satisfactory test of drive theory under these conditions, it would have been necessary to have a second transfer condition to serve as a control. This second condition would be one in which an entirely new set of stimulus cards containing new dimensions and values was used on the transfer problem. In this situation, the subject would be aware of a new problem beginning, and it would be less likely that the previously learned responses would compete with performance. The test of drive theory would be to determine whether the relationship between state anxiety and performance would be relatively more debilitating on the first transfer condition than on this second transfer condition. Such an investigation is currently in progress.

It is important to consider the present investigation as one of a series of studies attempting to clarify past research that has inconsistently reported debilitating effects of anxiety on concept learning. The thrust of this series of research is to demonstrate that a consideration of cognitive aptitudes and state anxiety would help to produce more consistent results. Presently, two of these studies have been completed using the same concept-learning task, and as expected, when trait anxiety has been considered by

itself, there have been no significant relationships to concept-learning performance. On the other hand, when trait anxiety was considered in conjunction with several aptitude measures (Meyers & Dunham, 1972) and when state anxiety was considered in the present investigation, significant debilitating relationships between anxiety and concept-learning performance have been found. Consequently, there is preliminary evidence suggesting the importance of considering aptitudes and state anxiety in research investigating the relationships between anxiety and concept learning.

REFERENCES

Denny, J. P. Effects of anxiety and intelligence on concept formation. *Journal of Experimental Psychology,* 1966, **72,** 596-602.

Maltzman, I., Fox, J., & Morrisett, L., Jr. Some effects of manifest anxiety on mental set. *Journal of Experimental Psychology,* 1953, **46,** 50-54.

Martin, R. P. Development of anxiety and coping processes in persons anticipating a high stress situation. Unpublished doctoral dissertation, University of Texas, Austin, 1970.

Martin, R. P., & Meyers, J. Effects of anxiety on quantity of examination preparation. Paper presented at the meeting of the American Educational Research Association, Chicago, April 1972.

Meyers, J., & Dunham, J. L. Anxiety-aptitude interactions in concept learning. *Proceedings of the 80th Annual Convention of the American Psychological Association,* 1972, **7,** 485-486.

O'Neill, H. F., Jr., Spielberger, C. D., & Hansen, D. H. Effects of state anxiety and task difficulty on computer-assigned learning. *Journal of Educational Psychology,* 1969, **60,** 343-350.

Sarason, I. G. Effect of anxiety, motivational instructions, and failure on serial learning. *Journal of Experimental Psychology,* 1956, **51,** 253-260.

Sarason, I. G., & Palola, E. G. The relationship of test and general anxiety, difficulty of task, and experimental instructions to performance. *Journal of Experimental Psychology,* 1960, **59,** 185-191.

Spence, J. T., & Spence, K. W. The motivational components of manifest anxiety: Drive and drive stimuli. In C. D. Spielberger (Ed.), *Anxiety and behavior.* New York: Academic Press, 1966.

Spielberger, C. D. (Ed.) *Anxiety and behavior.* New York: Academic Press, 1966.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *Manual for the State-Trait Anxiety Inventory.* Palo Alto, Calif.: Consulting Psychologist Press, 1970.

(Received October 19, 1972)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 40–48

EFFECTS OF POSITION AND TYPE OF QUESTION ON LEARNING FROM PROSE MATERIAL: INTERACTION OF TREATMENTS WITH INDIVIDUAL DIFFERENCES

RICHARD J. SHAVELSON[1]

Stanford University

DAVID C. BERLINER

Far West Laboratory for Educational Research and Development San Francisco, California

MICHAEL M. RAVITCH and DAVID LOEDING

Stanford University

Subjects were randomly assigned to one of four treatment groups: *position* (before or after) and *type* (lower order vs. higher order) of question placed in prose material. A control group had no questions in text. Before instruction, subjects received five aptitude tests. Instruction consisted of a 1,525-word prose passage. Immediately after and two weeks following instruction, subjects were tested. Differences in group means on four measures of achievement usually were statistically significant but small. A vocabulary test interacted with treatments. Subjects with low vocabulary scores might be assigned to text material with higher order questions placed after a prose passage, while subjects with high vocabulary scores might be assigned to text without questions.

The concept of mathemagenic activities (Rothkopf, 1965, 1970) has focused attention on intended and incidental learning as a consequence of adjunct questions inserted within prose material. The effect of question *position* (before vs. after prose passages) has been examined thoroughly. In general, a consistent facilitative learning effect occurs when questions are inserted after rather than before prose passages (Frase, 1968; Rothkopf, 1966; Rothkopf & Bisbicos, 1967). More recently the effects of the *type* of question (e.g., application, synthesis, or higher order question vs. factual or lower order question) have been explored (Allen, 1970; Hunkins, 1968; Tenenberg, 1969; Watts & Anderson, 1971). For example, Hunkins' data support the hypothesis that higher order questions prompt more thorough study and cognitive reorganization of the material, while lower order questions influence attention to facts. Similarly, Watts and Anderson found that application questions, as contrasted to questions that repeated examples or questions on names of significant people, yield the greatest general facilitative effect on learning by prompting the student to inspect and comprehend the text more thoroughly. Carroll (1971), in a review of this area, speculated that "questions are most effective when they not only cause memory search, but also cause some sort of reorganization of memory traces and associations [p. 164]."

When mean differences between treatment groups are examined, the effects of position and type of question are statistically significant, though usually small. But suppose that for particular subjects with similar scores on personality or ability measures, the effects of these variables are quite different from the effects for some other group of

[1] Requests for reprints should be sent to Richard J. Shavelson, who is now at the Graduate School of Education, University of California, Los Angeles, California 90024.

subjects in the treatment. Such effects cannot be detected when a mean is calculated for all subjects in the treatment. A certain combination of treatment variables, say higher order questions inserted after a passage, may result in a facilitative learning effect for a particular subgroup of subjects. For different subjects these treatment variables may not be facilitative, though such an effect could occur in some other treatment. These are hypotheses about aptitude by treatment interactions, or simply, ATI (see Berliner & Cahen, 1973; Cronbach & Snow, 1969).

Two ATI studies that are related to the present investigation have been completed. Berliner (1971) examined the effects of factual questions placed at specified intervals in a lecture presentation. In many analyses, for subjects with low scores in memory ability, questions inserted after lecture segments facilitated learning as measured by short-answer immediate- and delayed-retention tests. For subjects with high scores in memory ability, the questions may have interfered with learning. In a prose-learning study, Hollen (1971) found disordinal interactions between associative memory and questioning treatments. When the treatment did not contain adjunct questions, the need for associative memory was maximized (i.e., there was a positive slope for the regression of a learning measure on memory ability). However, when adjunct questions constituted the treatment, the need for associative memory was minimized (i.e., there was a negative slope for the regression of a learning measure on memory ability). To optimize treatment effects, then, subjects with low scores on memory ability might be assigned to the adjunct-question treatment and subjects with high scores might be assigned to a question-free treatment.

The Berliner and Hollen studies raise questions about the main effects of adjunct questions inserted in instructional material. The purpose of this study is to extend the examination of interactions between a subject's aptitudes and instructional treatments that differ in position and type of questions inserted in prose material.

Method

Subjects

Eighty-seven volunteer subjects from a local junior college were placed at random into one of five groups: (*a*) LB = lower order questions before text ($n = 18$); (*b*) LA = lower order questions after text ($n = 21$); (*c*) HB = higher order questions before text ($n = 20$); (*d*) HA = higher order questions after text ($n = 13$); and (*e*) C = control, no questions in text ($n = 15$). The proportion of male subjects ranged between .22 in Group LB and .45 in Groups HB and C. The mean age ranged between 21.3 years in Group LB and 24.2 years in Group C.

Materials

The instructional material, entitled "The Lisbon Earthquake," reported on the earthquake in Lisbon, Portugal, in 1755 and the historical and philosophical events that accompanied the quake. The material was selected for its novelty, reading level, reading time, reported interest, and its corresponding test items. Kropp, Stoker, and Bashaw (1966) constructed test items on "The Lisbon Earthquake" and had high school subjects read the text and answer the questions. Their data analysis showed that, in general, the empirical classification of items agreed with an a priori classification based on Bloom (1956). From the empirically determined taxonomic classification of test items, the questions for this study were selected. Lower order questions required subjects to demonstrate knowledge: "The size of the tidal wave which hit the Lisbon Harbor Area was: (a) 30 feet; (b) 40 feet; (c) 50 feet; (d) 60 feet." Higher order questions required subjects to demonstrate comprehension, application, and analysis (cf. Bloom, 1956): "A 'mental seismograph' is a (a) scientific device for detecting ideas; (b) figure of speech for the mind; (c) mental record; (d) mechanical device for recording earthquakes."

The 1,525-word text was divided into eight sections, the first seven consisting of two paragraphs each and the last section of four short paragraphs. For the four experimental groups, a lower or higher order multiple-choice question was inserted either before or after each section of text. Thus, a total of eight questions was inserted in the text for each experimental group. For Groups LB and HB, questions were inserted before each section of text and repeated, with the correct answers, at the end of the section. For Groups LA and HA, questions were placed after each section of text and repeated, with the correct answer, on the following page. The control group read the text without inserted questions and answers.

Questions were assigned to the text and the achievement test in the following way. A pool of 32 multiple-choice questions was created by including 2 lower order and 2 higher order questions from each of the text's eight sections. For the

lower order question treatment, 8 lower order questions were randomly assigned to both the text and the achievement test. The remaining 8 lower order questions appeared on the achievement test only. The 16 higher order questions were distributed in the same way. The achievement test, then, used all 32 items. It contained 8 lower order questions that appeared in the text, 8 higher order questions that appeared in the text, and 16 questions that were not used in the text.

Instruments

Five aptitude measures were administered. Three of these tests were from the battery collected by French, Ekstrom, and Price (1963). The Advanced Vocabulary Test (Part 2) measures verbal comprehension. It is a test of the ability to understand the English language. The Hidden Figures Test (Part 1) measures "the ability to keep one or more definite configurations in mind so as to make identifications in spite of perceptual distractions [p. 9]." The Advanced Vocabulary and Hidden Figures Tests may be considered tests of different general ability (intelligence) factors. Since different treatments placed differential emphasis on verbal comprehension and the ability to hold ideas in mind during instruction, these tests might be expected to interact with the treatments. The third test, Letter Span (Part 1) measures "the ability to recall perfectly for immediate reproduction a series of items after only one presentation of the series [p. 26]." This test was chosen because the ability measured had previously been shown to interact with questioning treatments (Berliner, 1971; Hollen, 1971).

The fourth test was the Taylor Manifest Anxiety Scale (Taylor, 1951). In this study, the test was labeled "Biographical Inventory." It positions subjects on an anxious to nonanxious continuum. It was included because a pilot study showed that high-anxious subjects performed better on a learning measure when questions preceded text material, whereas low-anxious subjects performed better when questions followed text material. Presumably the structure provided by presenting the questions before reading a passage attenuated their anxiety and permitted them to concentrate on learning the material.

The fifth test, memory for semantic implications, was constructed specifically for this study. This test was designed to reflect Guilford's (1967) description of an ability in his structure-of-intellect model. The test purports to measure the ability to remember *and* transform information presented in written material. The need for such a test was based on a psychological analysis of the cognitive requirements for learning prose material with adjunct questions. An answer to a higher order question seemed to require the ability to remember and transform the information in the text. This theorizing corresponds to Carroll's (1971) hypothesis about the effects of cognitive reorganization noted above. The reliability (Kuder-Richardson Formula 21) of the specially constructed test was .85.

The achievement test has already been described. Most of the items were taken from Kropp et al. (1966). Eight new lower order questions were constructed because of the need to have equal numbers of questions corresponding to each passage in text. The correlation between achievement test scores for immediate and retention testing, a rough index of test–retest reliability, was .71.

Procedure

The experiment was conducted over three one-hour sessions. In the first session, the experimenter explained the sequence of the study, assured confidentiality of test results, prompted the subjects to do their best, and explained that the purpose of the study was "to investigate how people learn, particularly how they learn from written materials." Following the introduction, test packages containing the five aptitude tests were distributed. Subjects were instructed to write their name, age, and sex on the package. The testing sequence was: Advanced Vocabulary (4 minutes); Biographical Inventory (7½ minutes); Hidden Figures (10 minutes); a rest break (4 minutes); Letter Span (5 minutes); and memory for semantic implications (7 minutes).

The second session was conducted one week later. Instructional materials corresponding to the LB, LA, HB, HA, or C conditions had been randomly ordered and were distributed to subjects at this time. Upon completing the study of the materials, subjects took the achievement test (immediate posttest).

The third session was conducted two weeks after the second session. The achievement test was administered again (retention testing).

After all data were collected and the preliminary analysis was made, the experimenter returned to the junior college and discussed the study with interested subjects.

Results and Discussion

Measures for Examining Learning

The achievement test contained 32 items classified into four groups: (*a*) lower order questions that appeared in text, (*b*) higher order questions that appeared in text, (*c*) lower order questions that did not appear in text, and (*d*) higher order questions that did not appear in text. In addition to a *total* score, scores on the four groups of questions can be combined in several ways to investigate learning. For example, if scores on lower and higher order questions that did not appear in text ("no text") are combined, a measure of *incidental* learning is

formed. If scores are calculated for lower order questions/no text and higher order questions/no text separately, measures of *transfer* from type of question in text to the same type of question are formed for lower and higher order question groups, respectively. Finally, if scores on test questions that also appear in text are calculated, a measure of *intentional* learning is obtained.

Intercorrelations among these measures are presented in Table 1, (higher order groups) and Table 2 (lower order groups). Most are positive and significantly different from zero. The highest are part–whole correlations, in which the items on one measure constitute a portion on another. For example, half of the items forming the total score constitute the incidental-learning score. The lowest intercorrelations are between transfer and intentional scores. These measures are independent of one another. Further, some subjects may be at ceiling on the intentional measure (see Table 3), and this restriction of range may be responsible for the lower correlations. Finally, note the difference in the correlations between intended and incidental measures in the two tables. This suggests that practice on and/or inspection of text for lower order questions does not facilitate answering questions on incidental material (see Table 2). This is not the case when higher order questions are used in text (see Table 1); they seem to facilitate learning of incidental material.

Effect of Placement and Type of Question on Learning: Comparison of Means

In Table 3, means[2] for each measure are reported by treatment group and time of testing (immediate and retention). Data on the number of subjects within each group and the standard deviations, especially for the total score, indicate that the assumption of homogeneity of variance in statistical tests will be violated. Thus, although all the tests are performed at $\alpha = .05$, the exact level of significance cannot be specified. Therefore, the tests of significance and the Johnson-Neyman (1936) analyses reported below should be interpreted with caution. Also, in these analyses, all comparisons between means are more conservative than would be the case if sample sizes were the same for each treatment.

[2] Additional descriptive data such as standard deviations and numbers of subjects are available from the first author.

TABLE 1

INTERCORRELATIONS BETWEEN MEASURES: HIGHER ORDER GROUPS

Score	1	2	3	4
1. Total		.90*	.73*	.59*
2. Incidental	.91*		.82*	.33*
3. Transfer	.71*	.88*		.23
4. Intended	.65*	.48*	.24	

* $p < .05$.

Total score. Total score data were examined with a Treatment Group × Time of Testing (5 × 2) analysis of variance with repeated measures on the time factor. Means are presented in Table 3. The group effect was significant ($F = 3.37$, $df = 4/65$). Post hoc comparisons of group means using the Newman-Keuls method for unequal *n*s showed that Group HA scored significantly higher than Group LB. Other pairwise comparisons showed no reliable differences. The time effect, as anticipated, was significant ($F = 28.48$, $df = 1/65$). Scores were higher at immediate testing ($\bar{X} = 23.04$) than at retention testing ($\bar{X} = 21.06$). The Group × Time interaction was not significant ($F < 1.0$). Orthogonal contrasts were performed to test the effects of (*a*) type of question (lower order vs. higher order), (*b*) position of question in text (before vs. after), and (*c*) their interaction. Reliable differences were not found. Nevertheless, the means as presented in Table 3 show a generally

TABLE 2

INTERCORRELATIONS BETWEEN MEASURES: LOWER ORDER GROUPS

Score	1	2	3	4
1. Total		.88*	.79*	.19
2. Incidental	.90*		.89*	.02
3. Transfer	.73*	.84*		.06
4. Intended	.51*	.24	.12	

* $p < .05$.

TABLE 3
MEAN SCORES FOR ALL MEASURES

Measure (maximum score)	Lower order questions before text	Higher order questions before text	Lower order questions after text	Higher order questions after text	Control
Total score (32)					
Immediate	20.69	23.50	23.00	25.31	23.00
Retention	19.59	21.56	20.06	24.30	21.00
Incidental learning (16)					
Immediate	9.50	11.05	9.75	11.31	11.36
Retention	9.12	10.11	8.50	11.40	10.69
Transfer: Lower order/no text (8)					
Immediate	5.31	5.85	5.60	6.31	6.21
Retention	4.94	5.22	4.69	6.30	5.70
Transfer: Higher order/no text (8)					
Immediate	4.06	5.20	4.15	5.00	5.14
Retention	4.18	4.89	3.81	5.00	5.00
Intentional learning (8)[a]					
Immediate	7.38	7.40	7.15	7.00	5.83
Retention	5.59	6.31	6.28	6.80	5.16

[a] For the four question groups, intended learning is calculated from scores on questions inserted in text. For example, the score for the group lower order questions before text is calculated from scores on low-order items appearing in text. Since the control group did not receive questions in text, the mean for both high- and low-order questions combined is reported as a baseline for comparison.

facilitative effect for higher order questions rather than for lower order questions and for questions after rather than questions before. These data, though not significant, conform to the findings in the literature reviewed above.

Incidental-learning measure. Incidental learning scores were examined with a Treatment Group × Time of Testing (5 × 2) analysis of variance with repeated measures on the time factor. Means are presented in Table 3. The group effect was significant ($F = 3.00$, $df = 4/65$). Pairwise comparisons showed no reliable differences. The time effect was significant ($F = 13.49$, $df = 1/65$). Scores were higher at immediate testing ($\bar{X} = 9.81$). The Group × Time interaction was not significant ($F < 1.0$). Orthogonal contrasts were performed to test the effects of (*a*) type of question, (*b*) position of question, and (*c*) their interaction. Reliable differences were not found. The trends among the means in this analysis conform to the findings on question position and type.

Transfer-learning measure. Means for the transfer scores are presented in Table 3. Two separate Treatment Group × Time of Testing (3 × 2) analyses of variance with repeated measures on the time factor were performed on transfer scores for the lower order plus control and higher order plus control groups. The only main effect found was time of testing for the lower order groups plus control ($F = 8.66$, $df = 1/40$). A transfer effect from a certain type of question placed in text to the same type of question never before seen does not occur.

Intentional-learning measure. Means for the intentional scores are presented in Table

3. Two separate Treatment Group × Time of Testing (3 × 2) analyses of variance with repeated measures on the time factor were performed on the intentional-learning scores for the lower order plus control and higher order plus control groups. In both analyses, a significant group main effect was obtained ($F = 3.52$, $df = 2/40$; $F = 6.05$, $df = 2/36$; respectively). As expected, exposure to questions placed in text produces performance that is near ceiling on the same items on the criterion test.

In the discussion of the results from the four learning measures, the performance of the control group cannot be overlooked. Except for Group HA, Group C usually performed as well as or better than the other experimental groups. Although an explanation for this finding is not readily available, it may involve the short length of the prose material.

Aptitude by treatment interactions. If the five treatment conditions have similar effects on all individuals and if achievement test scores are regressed on aptitude scores, the regression slopes for treatments should be parallel and the difference among slopes can be explained by the differences in means among the groups. But if the treatment groups do not have similar effects on all individuals (i.e., persons high on Aptitude A do well with higher order questions placed after text, but not with these questions placed before text), the regression slopes for treatments should not be parallel. Rather, they should interact and perhaps cross at some point. To determine if aptitude by treatment interactions were present, total- and incidental-learning scores were regressed on aptitude scores. Transfer- and intentional-learning measures were not examined due to the restricted range (0–8 points) and ceiling effect for these data. Separate analyses of variance of scores from each aptitude test showed that the groups did not differ significantly on these variables. Except for the correlations between the Advanced Vocabulary Test and the Biographical Inventory ($r = -.22$) and the Hidden Figures Test and the memory for semantic implications ($r = .23$), intercorrelations among aptitude tests were not significantly different from zero.

A computer program (Dowaliby & Berliner, 1971) using the Johnson-Neyman (1936) technique with the Potthoff (1964) modification tested the data for interactions. For each set of analyses, the hypothesis of a common slope (achievement regressed on aptitude for every possible pairwise combination of treatment groups) is tested ($\alpha = .05$). For significant interactions, the Johnson-Neyman technique is applied such that a region of nonsignificance is determined ($\alpha = .10$). At this point the alpha level should be determined by the decision maker, since it depends on the type of risk he is willing to take in classifying students. Cases falling within the region of nonsignificance may be assigned to either treatment; cases falling outside this region should be assigned to one or another treatment.

This analysis did not support the hypothesized interaction between the memory ability tests (Letter Span and memory for semantic implications) and treatments. Also, the Biographical Inventory and the Hidden Figures Test did not interact with treatments.[3]

Total score. Advanced Vocabulary Test scores interacted with treatment at immediate and retention testing (see Figures 1 and 2). A disordinal interaction appeared with immediate-testing data. The regression slope for Group HA differs significantly from the slope for Groups LA and C. Subjects in Groups LA and HA with Advanced Vocabulary scores above 8.60 may be assigned to either treatment. Subjects with scores below 8.60 on Advanced Vocabulary should be assigned to the HA treatment. Fifty-six percent of the subjects in Groups LA and HA should be assigned to Group HA to optimize their achievement. Subjects in Groups HA and C with Advanced Vocabulary scores above 8.09 may be assigned to either treatment. Subjects with scores below 8.09 should be assigned to the HA treatment. Sixty-seven percent of the subjects in Groups HA

[3] The ATI data on these measures are available from the first author.

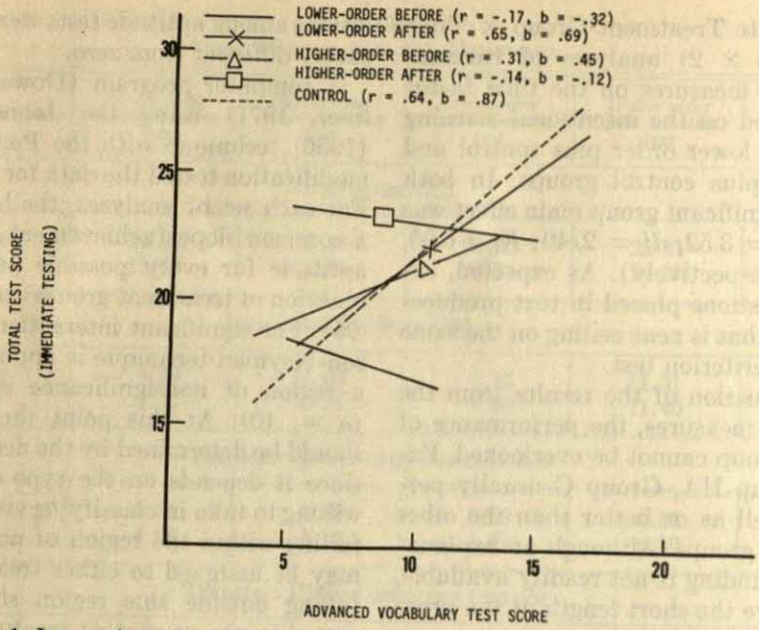


FIGURE 1. Interaction of advanced vocabulary scores with treatment effects as measured by the achievement total score at immediate testing.

and C should be assigned to the HA treatment.

For retention-testing data, the regression slope for Group HA differs significantly from the slope for Group LA. Subjects with scores above 8.98 on Advanced Vocabulary may be assigned to either treatment. Subjects with scores below 8.98 should be as-

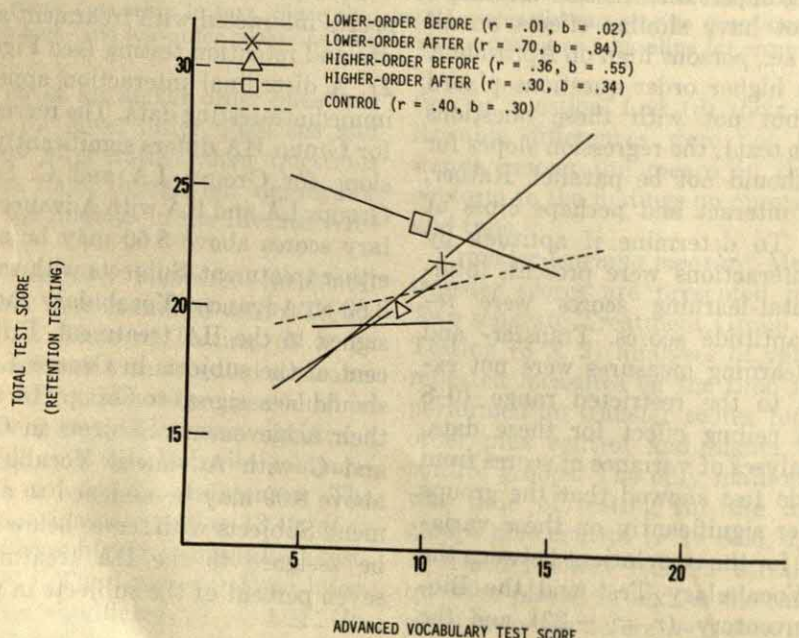


FIGURE 2. Interaction of advanced vocabulary scores with treatment effects as measured by the achievement total score at retention testing.

signed to the HA treatment. Fifty-seven percent of the subjects in Groups HA and LA should be assigned to the HA treatment.

Incidental-learning measure. Advanced Vocabulary also interacts with treatment using the incidental-learning measure taken at immediate and retention testing. At immediate testing, subjects in Groups LA and and HA ($\beta = .45$ vs. $\beta = -.09$) with aptitude scores below 5.27 should be assigned to the HA treatment; subjects with scores above this value may be assigned to either treatment. Nine percent of the subjects in Groups LA and HA should be assigned to the HA treatment. Subjects in Groups HA and C ($\beta = -.09$ vs. $\beta = .47$) with Advanced Vocabulary scores below 3.37 should be assigned to HA; subjects with scores above 17.63 should be assigned to the C treatment. None of the subjects in these groups had scores beyond the critical values.

At retention testing, subjects in Groups LA and HA ($\beta = .57$ vs. $\beta = -.15$) with Advanced Vocabulary scores below 8.50 should be assigned to HA; subjects with scores above that value may be assigned to either treatment. Fifty-seven percent of the subjects in Groups LA and HA should be assigned to the HA treatment.

The Advanced Vocabulary Test is related to measures of general mental ability that, in most studies on instruction, yield consistent positive correlations with outcome measures. The unique finding in this study is a slightly negative correlation between scores on the Advanced Vocabulary Test and the outcome measure for subjects in the HA treatment. In conjunction with positive correlations for Groups LA and C, disordinal interactions were obtained.

The insertion of HA questions appears to aid subjects with low Advanced Vocabulary scores. Questions of this type and in this position may act in a compensatory manner (Salomon, 1972) when the ability to interrelate concepts and ideas in the prose material is deficient. The HA questions prompt subjects to link together concepts in the prose material and/or to link the concepts in the prose material to the subjects own existing cognitive structure. The effectiveness of the HA question may also lie in its ability to promote review in subjects who ordinarily do not engage in such activities.

The negative slope also indicates that the HA treatment interferes with prose learning for subjects high in verbal comprehension. These subjects may possess effective strategies for processing prose material that are disrupted by insertion of external prompts.

If upon replication this finding is confirmed, there are immediate actions that can be taken to tailor instruction to the needs of particular learners. Verbal ability measures exist or are readily obtained in most instructional settings. Instructional material for those who score low on such measures can be modified to include the insertion of higher order questions. If this technique can compensate for certain deficiencies in study skills, a large number of students can be helped in learning from prose material.

REFERENCES

ALLEN, D. I. Some effects of advanced organizers and level of question on the learning and retention of written social studies material. *Journal of Educational Psychology*, 1970, **61**, 333–339.

BERLINER, D. C. Aptitude-treatment interactions in two studies of learning from lecture instruction. Paper presented at the meeting of the American Educational Research Association, New York, February 1971.

BERLINER, D. C., & CAHEN, L. S. Trait-treatment interactions and learning. In F. N. Kerlinger (Ed.), *Review of research in education*. Itasca, Ill.: Peacock, 1973.

BLOOM, B. S. (Ed.) *Taxonomy of educational objectives, Handbook I: Cognitive domain*. New York: McKay, 1956.

CARROLL, J. B. *Learning from verbal discourse in educational media: A review of the literature*. (Research Bulletin RB-71-61) Princeton, N. J.: Educational Testing Service, October 1971.

CRONBACH, L. J., & SNOW, R. E. *Individual differences in learning ability as a function of instructional variables*. (Final report, U. S. Office of Education Contract OEC-4-6-061269-1217) Stanford, Calif.: Stanford University, School of Education, March 1969.

DOWALIBY, F. J., & BERLINER, D. C. *ANALATI: A computer program for the application of the Johnson-Neyman technique in the case of one predictor and one criterion variable*. (Tech. Rep. 15) Amherst: University of Massachusetts, School of Education, Center for Educational Research, November 1971.

FRASE, L. T. Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology*, 1968, **59**, 244–249.

FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. *Manual for Kit of Reference Tests for Cognitive Factors.* Princeton, N. J.: Educational Testing Service, 1963.

GUILFORD, J. P. *The nature of human intelligence.* New York: McGraw-Hill, 1967.

HOLLEN, T. T., JR. Interaction of individual abilities with the presence and position of adjunct questions in learning from prose materials. *Dissertation Abstracts International,* 1971, **31,** 5847A.

HUNKINS, F. P. The effects of analysis and evaluation questions on various levels of achievement. Paper presented at the meeting of the American Educational Research Association, Chicago, February 1968.

JOHNSON, P. O., & NEYMAN, J. Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs,* 1936, **1,** 57–93.

KROPP, F. F., STOKER, H. W., & BASHAW, W. L. The construction and validation of tests of the cognitive processes as described in the "Taxonomy of Educational Objectives." (U.S. Office of Education Contract OE-4-10-019) Tallahassee: Florida State University, February 1966.

POTTHOFF, R. F. On the Johnson-Neyman technique and some extensions thereof. *Psychometrika,* 1964, **29,** 241–256.

ROTHKOPF, E. Z. Some theoretical and experimental approaches to problems in written instruction. In J. D. Krumboltz (Ed.), *Learning and the educational process.* Chicago: Rand-McNally, 1965.

ROTHKOPF, E. Z. Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal,* 1966, **3,** 241–249.

ROTHKOPF, E. Z. The concept of mathemagenic activities. *Review of Educational Research,* 1970, **40,** 325–336.

ROTHKOPF, E. Z., & BISBICOS, E. E. Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology,* 1967, **58,** 56–61.

SALOMON, G. Heuristic models for the generation of aptitude-treatment interaction hypotheses. *Review of Educational Research,* 1972, **42,** 327–343.

TAYLOR, J. A. The relationship of anxiety to the conditioned eyelid response. *Journal of Experimental Psychology,* 1951, **41,** 81–92.

TENENBERG, M. S. Effects of different study instructions on learning from written material. Unpublished doctoral dissertation, University of California, Berkeley, 1969.

WATTS, G. H., & ANDERSON, R. C. Effects of three types of inserted questions on learning from prose. *Journal of Educational Psychology,* 1971, **62,** 387–394.

(Received November 2, 1972)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 49-51

NOTE ON CRANO, KENNY, AND CAMPBELL'S "DOES INTELLIGENCE CAUSE ACHIEVEMENT?"

JEAN L. DYER[1] AND LOUISE B. MILLER

University of Louisville

The cross-lagged panel correlation technique was applied to achievement and intelligence data on young inner-city children. Findings indicated the predominant causal sequence was that of achievement causing intelligence as suggested by the Crano, Kenny, and Campbell data.

Recently, Crano, Kenny, and Campbell (1972) used the cross-lagged panel correlation technique to investigate the direction of the causal relationship between intelligence and achievement. Such an analysis requires correlational information on two variables at more than one point in time. Unlagged synchronous correlations and lagged autocorrelations are checked for significance and reliability, while the correlations that are crossed and lagged over time provide information regarding the direction of the causal relationship.

The Crano et al. data at the fourth- and sixth-grade levels on the Lorge-Thorndike intelligence test and the Iowa Test of Basic Skills generally supported the inference "that the predominant causal sequence is that of intelligence causing later achievement [p. 266]," as indicated by a significantly larger correlation between early intelligence and later achievement than between early achievement and later intelligence. However, comparison of inner-city and suburban school data indicated that the relationship was in the opposite direction for the inner-city schools, although this difference was not significant. The authors suggested that an abstract-to-concrete causal sequence predominated for the suburban children, while the opposite occurred for the inner-city children. They also suggested that similar studies conducted with younger children might provide a clearer picture of the achievement–intelligence relationship.

Intelligence and achievement data necessary for a cross-lagged panel analysis were available in a longitudinal comparison of the effects of four Head Start programs (Miller & Dyer, 1970). Such a group provided a further test of the Crano et al. hypothesis about inner-city children.

METHOD

Sample

A total of 213 children participated in the experimental Head Start programs during the Head Start year and were followed up in both the kindergarten and first-grade years. Achievement and intelligence tests were given in the spring of each year. Complete sets of achievement and intelligence data over the three years were available on 173 children. These children were all in the areas of the city eligible for Title I funding.

Tests

The Stanford-Binet intelligence test, Revision L-M, was individually administered to the children in the spring of each year. Testers were graduate students in psychology who had had testing practicums and previous experience in administering the Stanford-Binet to inner-city children. Internal consistency reliability coefficients for the Stanford-Binet, based upon coefficient α for stratified-parallel tests (Rajaratnam, Cronbach, & Gleser, 1965; Silverstein, 1969) were .90, .83, and .98, respectively, for the Head Start, kindergarten, and first-grade years.

The Preschool Inventory (Caldwell, 1968), a test of specific achievements representative of what the child brings to an educational situation rather than broad cognitive functioning, was administered in the spring of the Head Start and kindergarten

[1] Requests for reprints should be sent to Jean L. Dyer, School of Education, University of Louisville, Louisville, Kentucky 40208.

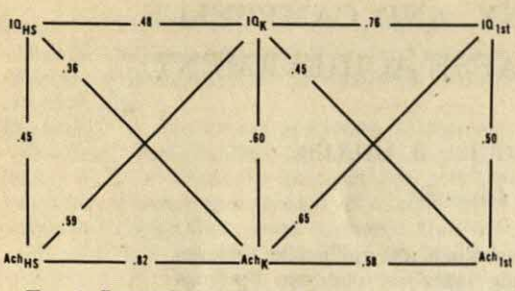


FIG. 1. Interrelationships between intelligence (IQ) and achievement (Ach) from Head Start (HS) through first grade at one-year intervals. (K = kindergarten; 1st = first grade.)

years. It focuses upon the vocabulary of the young child, his knowledge of basic numerical and sensory concepts, and his knowledge of his own personal world. For the Preschool Inventory, the Kuder-Richardson Formula 20 reliability coefficients were .89 and .85, respectively, for the Head Start and kindergarten years. First-grade achievement was measured by the California Achievement Test scores, which were obtained from public school records. Raw scores were available on both the reading and mathematics subtests. However, no internal consistency reliability measures were available. The Preschool Inventory was individually administered by testers trained by the project staff, while the California Achievement Test was administered in a group setting by the public school staff.

RESULTS

Since the Stanford-Binet and Preschool Inventory correlations with the California reading and mathematics subtest scores were identical, separate analyses on reading and mathematics were not made. The cross-lagged panel results of the correlations between intelligence and achievement for the two one-year intervals are shown in Figure 1. For a sample size of 173, all correlations greater than .12 were significant at the .05 level. The test–retest correlations (lagged, autocorrelations) were significant. The synchronous, unlagged correlations between achievement and intelligence were also significant.

It might be noted that relatively low test–retest correlations occurred between Head Start and kindergarten intelligence and between kindergarten and first-grade achievement. The latter may be due to the shift to a different achievement test at first grade. However, both Head Start and kindergarten intelligence were measured by the same test. It was hypothesized that the low correlation between intelligence measures at these two points might be attributed to the immediate influence of the experimental Head Start programs on the rank order of the students. In order to investigate this hypothesis, test–retest correlations between the beginning and end of Head Start measures for the experimental group referred to were compared with those for an available control group ($n = 34$) who had not been exposed to any Head Start program. On both achievement and intelligence, these test–retest correlations were lower for the experimental group (IQ $r = .70$; achievement $r = .79$) than for the control group (IQ $r = .83$; achievement $r = .94$).

Since reliable measurement of both achievement and intelligence was indicated by high internal consistency coefficients and significant test–retest correlations, comparison of the cross-lagged correlations was made. A significant difference occurred at both the Head Start – kindergarten interval ($t = 3.32$, $df = 170$, $p < .01$, two-tailed) and the kindergarten – first grade interval ($t = 3.24$, $df = 170$, $p < .01$, two-tailed).[2] The correlation between early achievement and later intelligence was significantly greater than that between early intelligence and later achievement. Comparison of the cross-lagged correlations over the two-year interval from Head Start to first grade ($r = .64$, between Head Start achievement and first-grade IQ, and $r = .38$, between Head Start IQ and first-grade achievement) produced a similar result ($t = 8.23$, $df = 170$, $p < .01$, two-tailed). Over this two-year period, the test–retest correlation for intelligence was .56 and for achievement was .41.

As mentioned previously, Crano et al. (1972) hypothesized from their results that for inner-city children the predominant sequence is that of achievement causing intelligence. The present findings provide stronger support for their hypothesis.

[2] The t test was the same as that used by Crano et al. (1972; see Peters & Van Voorhis, 1940, p. 185).

REFERENCES

CALDWELL, B. M. *Preschool Inventory: Experimental edition.* Princeton, N. J.: Educational Testing Service, 1968.

CRANO, W. D., KENNY, D. A., & CAMPBELL, D. T. Does intelligence cause achievement?: A cross-lagged panel analysis. *Journal of Educational Psychology,* 1972, **63,** 258–275.

MILLER, L. B., & DYER, J. L. Experimental variation of Head Start curricula: A comparison of current approaches. *ERIC,* 1970, No. ED 045 196.

PETERS, C. C., & VAN VOORHIS, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill, 1940.

RAJARATNAM, N., CRONBACH, L. J., & GLESER, G. Generalizability of stratified-parallel tests. *Psychometrika,* 1965, **30,** 39–56.

SILVERSTEIN, A. B. The internal consistency of the Stanford-Binet. *American Journal of Mental Deficiency,* 1969, **73,** 753–754.

(Received December 11, 1972)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 52–56

INITIAL LEVEL OF STUDENT EVALUATION OF INSTRUCTION AS A SOURCE OF INFLUENCE ON INSTRUCTOR CHANGE AFTER FEEDBACK[1]

HAGOP S. PAMBOOKIAN[2]

University of Michigan

In early October 1971, 252 students in 13 introductory and educational psychology sections responded to a Student Opinion Questionnaire containing measures of seven stable dimensions on college teaching. Ten days later, the instructors, who had been grouped according to the initial level of student evaluation, received feedback. In December, 231 students responded again to the Student Opinion Questionnaire. The students' initial evaluation of instruction was found to be a significant source of influence on instructor change after feedback. Instructors who were evaluated moderately well improved their teaching significantly on skill, interaction, and rapport, when compared to instructors rated more favorably. They showed trends toward decreasing work overload and improving rapport when compared to instructors rated more unfavorably.

Instructors can improve their teaching through various approaches, and one way of helping them bring positive changes in their classroom behavior is to tell them what their students think of their teaching.

Not much has been done with feedback in college classrooms. Most of the research on feedback in teaching has been carried out with elementary and secondary school teachers (e.g., Gage, Runkel, & Chatterjee, 1963; Salomon & McDonald, 1970; Shively, Van Mondfrans, & Reed, 1970; Tuckman & Oliver, 1968). Several investigators showed that when student evaluations of teachers were fed back to teachers, they improved their teaching in the sixth grade (Gage et al., 1963), in Grades 7–12 (Bryan, 1963), and in high school or a technical institute (Lauroesch, Pereira, & Ryan, 1969; Tuckman & Oliver, 1968). Daw and Gage (1967), likewise, found that feedback from elementary school teachers improved their principals' behavior; while Tuckman and Oliver (1968) showed that the source of feedback had different effects on teachers, that is, student feedback improved teachers' behavior but feedback from supervisors alone had an adverse effect on them.

Other studies using feedback, however, reported no significant changes in vocational agricultural teachers (Thomas, 1969), in graduate teaching assistants in religion and earth science (Miller, 1971), and in teaching fellows teaching introductory and educational psychology courses (Pambookian, 1972).

The present study was concerned with college students' evaluations of instruction and change by the instructor after learning of such ratings. It was believed that the students' initial evaluations of instruction significantly influence the extent to which instructors change their teaching as a result of feedback. The degree of students' satisfaction and dissatisfaction with teaching,

[1] This article is based on portions of a dissertation submitted in partial fulfillment of the requirements for the doctoral degree at the University of Michigan.

The author wishes to thank Wilbert J. McKeachie, Allen Menlo, Alvin F. Zander, and David Starks for their assistance during all phases of the study. A slightly different version of this paper was presented at the annual meeting of the Midwestern Psychological Association, Chicago, May 1973.

[2] Requests for reprints should be sent to Hagop S. Pambookian, 541 S. Ashley Street, Ann Arbor, Michigan 48103.

therefore, was thought to be an important factor in the use of feedback to improve teaching. It was hypothesized that instructors evaluated moderately well by their students prior to feedback change more after feedback than instructors evaluated more favorably or more unfavorably.

It is likely that instructors who are given favorable evaluations of their teaching by their students are satisfied with their behavior in the classroom. Hence, they may see no need for change and/or improvement. These instructors have been positively reinforced in their teaching and are likely to continue doing what was perceived to be skillful and effective. The instructors who have been evaluated more unfavorably by their students, on the other hand, may be made anxious by the ratings. A high-anxiety state is likely to lead to rigidity in behavior that may have disruptive and interfering effects on teaching. The instructors who have been evaluated moderately well, however, are likely to be more responsive to information (i.e., feedback) from students and more eager to improve their teaching. Unlike the others, these instructors are neither elated by more favorable evaluations nor disillusioned by more unfavorable ratings by students and become more attentive to their perceptions and needs, thus bringing about positive changes in their behavior and teaching. For them, an improvement in teaching behavior is a feasible but challenging goal.

Method

Subjects

Thirteen teaching fellows (hereafter referred to as instructors) teaching sections of psychology at the University of Michigan participated in this study. All of them were advanced doctoral students in various psychology programs. Eight of the instructors were teaching a multisection introductory psychology course, while the remaining five were teaching an introductory course in educational psychology. In both cases, the instructors had freedom in planning their sections the way they and/or their students wanted, emphasizing certain aspects of the course and deemphasizing others based on their background and previous teaching experiences. There was flexibility as far as course objectives, content, procedures, assignments, and evaluations were concerned.

Measurement of Instructor Behavior

Behavior of the instructors was measured by the revised McKeachie-Lin Student Opinion Questionnaire originally developed by Isaacson, McKeachie, and Milholland (1963).[3] Using the 46-item questionnaire in introductory psychology courses at the University of Michigan, Isaacson, McKeachie, Milholland, Lin, Hofeller, Baerwaldt, and Zinn (1964) isolated six stable dimensions (or factors) of teaching on the college level as evaluated by students. They commented, "The factor similarity analysis suggests that six factors can be regarded as evidencing stability over sexes, evaluation periods, student groups, and teacher groups [p. 348]." However, based on the latest findings in Michigan studies, a seventh dimension was added: achievement standard.

The experimenter selected 21 items from the questionnaire.[4] Three items with high loadings for each of the seven dimensions were used in the study. (In addition, two items were included that assessed the instructor's *all-around teaching ability* and *the overall value of the course*.) The dimensions were (*a*) skill (instructor's ability in observing student reactions, explaining clearly, and stimulating the intellectual curiosity of the students); (*b*) overload (amount and degree of difficulty of work assigned by the instructor); (*c*) structure (the concern for keeping a tight schedule, following an outline closely, and deciding in detail what should be done and how); (*d*) feedback (keeping students well informed of their progress, telling them when they have done a particularly good job, or criticizing poor work); (*e*) interaction (students frequently volunteering their own opinions, arguing with one another or with the teacher, and expressing opinions and disagreements); (*f*) rapport (instructor's being friendly, permissive, flexible, and listening to what class members have to say); and (*g*) achievement standard (instructor's emphasis on grades, high-quality work, and maintenance of definite standards of student performance).

Feedback Procedure

The following three steps describe the sequence of data collection:

1. All the students took the Student Opinion Questionnaire during the fourth week of the fall 1971 term. Each instructor simultaneously responded to the questionnaire as to how he perceived his own teaching with regard to the items rated. In no instance were the students identified

[3] The latest available version of the questionnaire was given to the investigator by Yi-Guang Lin, a research psychologist at the University of Michigan, for use in this study.

[4] The items on the Student Opinion Questionnaire were rated on a 5-point scale. The points on the scale were: almost always, often, occasionally, seldom, and never.

by name or identification number. All of the participants were encouraged to make comments pertinent to the items and/or the questionnaire. Machine-scorable IBM answer sheets were used for the answers. (At this time, 252 students responded to the Student Opinion Questionnaire.)

2. A week or 10 days after the administration of the Student Opinion Questionnaire, the researcher spent from 30 to 45 minutes with each of the instructors, and the results of student evaluations were presented to him. The instructors were given feedback for (*a*) the Student Opinion Questionnaire by each item and teaching dimension, using the class means, as well as the frequencies and percentages for each option on the scale; (*b*) the Student Opinion Questionnaire ratings by student grade point average and sex; and (*c*) their own self-evaluations and the discrepancy scores between student evaluations of them and their own perceptions of their teaching.

It was made clear to the instructors that feedback was a means of providing them with information from their students relevant to their teaching and classroom behavior. They were permitted to ask questions regarding the results, but discussion concerning preferred changes in teaching style or ways of making these changes was not allowed. During the feedback sessions with the instructors, there seemed to be a genuine interest in the students' opinions. Their interest after seeing the results was expressed by "That's interesting"; "It's good to know that"; "Beautiful"; "I am surprised, really"; "I didn't think they were so understanding"; "I never thought they would respond that way"; "I must do something now"; "They must like me"; "I feel much better now!"

3. In early December, eight weeks after the feedback sessions, the Student Opinion Questionnaire was again administered to the students and instructors. The instructors also responded to a few questions related to student evaluations and the impact of feedback on them and their teaching. At that time, instructors, as well as students, were given an item entitled "Feedback to Participating Students," which extended appreciation for their cooperation. (Two hundred and thirty-one students responded to the Student Opinion Questionnaire.)

Measurement of Change in Instructor Behavior and Teaching

In order to have similar comparisons among the instructor groups on teaching dimensions, the ratings on each of the Student Opinion Questionnaire items were averaged across students on pre- and postfeedback means, and then the three item means for each dimension were added.

Instructors were grouped according to the initial ratings on evaluative items of skill and rapport (i.e., Dimensions *a* and *f*, described previously), as research (e.g., Isaacson et al., 1964; McKeachie, Lin, & Mann, 1971) has shown these dimensions to be clearly indicative of effective teaching. The three groups were (*a*) more favorably rated instructors ($n = 7$); (*b*) more moderately rated ($n = 3$); and (*c*) more unfavorably rated ($n = 3$). The groups thus formed were compared, however, on all seven teaching dimensions. Analysis of variance and *t* tests were employed to test whether the initial level of students' evaluations of instruction made any significant impact on instructor change and improvement after feedback.

Results and Discussion

Analyses of variance comparing changes in teaching behavior of instructors rated favorably, moderately, and unfavorably prior to feedback on each of the seven teaching dimensions, gave some evidence of differences among the groups on four dimensions. The differences were significant among the groups on rapport ($F = 4.11$, $df = 2/10$, $p < .05$), and there were strong trends toward differences on skill ($F = 3.23$, $df = 2/10$, $p < .08$), overload ($F = 3.58$, $df = 2/10$, $p < .07$), and interaction ($F = 3.24$, $df = 2/10$, $p < .08$).[5]

Individual *t* tests were conducted to compare the difference or gain scores between the instructor groups. The resulting *t* values indicated that the differences of gain scores between the favorably rated instructors and the moderately rated ones were significant in the teaching dimensions of skill ($t = 2.48$, $df = 10$, $p < .03$), interaction ($t = 2.52$, $df = 10$, $p < .03$), and rapport ($t = 2.86$, $df = 10$, $p < .02$).

It was evident from the postfeedback means of instructors in both groups that the instructors who had been evaluated moderately well prior to feedback changed (i.e., improved) their teaching significantly after the feedback in at least three dimensions compared to those who had been rated more favorably. After the feedback, the moderately rated instructors explained subject matter more clearly and their explanations were more often to the point; they became more skillful in observing student reactions; they more often stimulated the intellectual curiosity of their students; and they increased their interaction and rapport with the students. Based on the postfeedback scores, it can be concluded that the moder-

[5] In view of the small number of cases involved, analyses of variance and *t*-test tables are eliminated from inclusion in the article.

ately evaluated instructors brought more positive changes, and hence improvement, to their teaching than the initially favorably evaluated instructors. The latter tended to regress toward the mean. Figure 1 presents graphically the gains made by all the instructors on skill, interaction, and rapport.

The results also indicated that compared to the more unfavorably rated instructors, trends toward change were found in the moderately rated instructors in rapport ($t = 1.88, df = 10, p < .09$) and toward less work overload ($t = 1.90, df = 10, p < .09$) in teaching. In both cases, instructors evaluated moderately well seemed to benefit more from feedback than those more unfavorably rated.

It should be interesting to note also that comparing the moderately and favorably rated instructors on all-around teaching ability and the overall value of the course, the moderately rated instructors changed more and the changes were significantly positive on the overall value of the course ($t = 2.82, df = 10, p < .02$); they showed a strong trend toward significant difference with respect to all-around teaching ability ($t = 2.17, df = 10, p < .06$). It appears that feedback helped the moderately rated instructors to enhance the value of the course and to improve their general teaching ability.

Conclusion and Implications

The results give substantial support to confirm the hypothesis that moderately evaluated instructors change more after feedback, because they changed their teaching in several dimensions. It appears that feedback had a more positive effect on the teaching of instructors who were moderately evaluated prior to feedback than on other teachers.

As predicted, the initial level of student evaluation of instructor and instruction was found to be a strong source of influence on the instructor. The moderately evaluated instructors changed significantly in skill, interaction, and rapport compared to instructors who had been evaluated more favorably. The contrast in changes was significant at .05 level and in the expected direction.

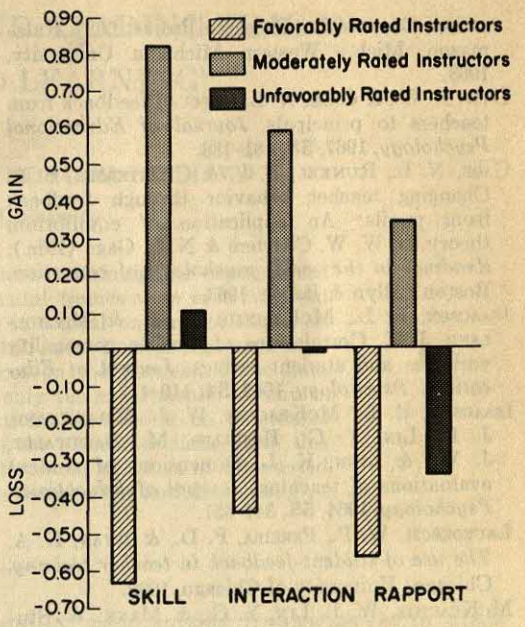


Fig. 1. Mean gain scores on each dimension for instructors evaluated more favorably, moderately well, and more unfavorably prior to feedback.

That is, after the feedback, the moderately evaluated instructors became more skillful in their teaching, had more classroom interactions and rapport as perceived by their students. (The differences between the moderately rated and the unfavorably rated instructors were not statistically significant.)

The implications for instruction are obvious: (*a*) In giving feedback to teachers in training and in service, the initial level of performance based on student evaluations should be taken into consideration; (*b*) teachers should be assisted in the development of skills to make better use of information on their teaching, which may not be very positive or favorable; and (*c*) other means should be explored for helping instructors whose teaching is poor or ineffective. *Feedback* as a means of bringing about positive change in, and improving, college instruction must be further investigated, however, with a larger number of faculty members in various departments to appraise its definitive role in teacher preparation and improvement.

REFERENCES

Bryan, R. C. *Reactions to teachers by students, parents, and administrators.* (U.S. Office of Edu-

cation, Cooperative Research Project 668). Kalamazoo, Mich.: Western Michigan University, 1963.

DAW, R. W., & GAGE, N. L. Effect of feedback from teachers to principals. *Journal of Educational Psychology,* 1967, **58,** 181–188.

GAGE, N. L., RUNKEL, P. J., & CHATTERJEE, B. B. Changing teacher behavior through feedback from pupils: An application of equilibrium theory. In W. W. Charters & N. L. Gage (Eds.), *Readings in the social psychology of education.* Boston: Allyn & Bacon, 1963.

ISAACSON, R. L., MCKEACHIE, W. J., & MILHOLLAND, J. E. Correlations of teacher personality variables and student ratings. *Journal of Educational Psychology,* 1963, **54,** 110–117.

ISAACSON, R. L., MCKEACHIE, W. J., MILHOLLAND, J. E., LIN, Y. G., HOFELLER, M., BAERWALDT, J. W., & ZINN, K. L. Dimensions of student evaluations of teaching. *Journal of Educational Psychology,* 1964, **55,** 344–351.

LAUROESCH, W. P., PEREIRA, P. D., & RYAN, K. A. *The use of student feedback in teacher training.* Chicago: University of Chicago, 1969.

MCKEACHIE, W. J., LIN, Y. G., & MANN, W. Student ratings of teacher effectiveness: Validity studies. *American Educational Research Journal,* 1971, **8,** 435–445.

MILLER, M. T. Instructor attitudes toward, and their use of, student ratings of teachers. *Journal of Educational Psychology,* 1971, **62,** 235–239.

PAMBOOKIAN, H. S. The effect of feedback from students to college instructors on their teaching behavior. Unpublished doctoral dissertation, University of Michigan, 1972.

SALOMON, G., & MCDONALD, F. J. Pretest and posttest reactions to self-viewing one's teaching performance on videotape. *Journal of Educational Psychology,* 1970, **61,** 280–286.

SHIVELY, J. E., VAN MONDFRANS, A. P., & REED, C. L. *The effect of mode of feedback in microteaching.* Washington, D.C.: U.S. Office of Education, 1970.

THOMAS, H. B., JR. Changing teacher behavior: An experiment in feedback from students to teachers. *Dissertation Abstracts International,* 1969, **5,** 1903–4A.

TUCKMAN, B. W., & OLIVER, W. F. Effectiveness of feedback to teachers as a function of source. *Journal of Educational Psychology,* 1968, **59,** 297–301.

(Received December 12, 1972)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 57-61

SEQUENCE EFFECTS AND READING TIME IN PROGRAMMED LEARNING[1]

JAMES W. DYER AND RAYMOND W. KULHAVY[2]

Arizona State University

Undergraduates read a programmed text containing one experimental and three placebo sections. Experimental frames were either massed or distributed over placebo sections in logical or random sequence. There were no effects for practice distribution. Logical sequences were recalled better than random sequences on the posttest. These data were replicated by groups receiving only the experimental section in either a logical or random presentation. There were no differences due to program length or presence of the placebo sections. Both this study and previous research were interpreted in terms of attenuated inspection behaviors due to both the distribution and disordering of frames.

The appropriate sequencing of text material has long been a canon of sound instructional design (cf. Gagné, 1965). Interestingly, studies comparing criterion achievement from logically sequenced teaching programs in which frames are randomly ordered yield inconsistent results (Niedermeyer, 1968). In spite of the careful attention paid to correct sequence, this research often leads to a no-difference verdict (e.g., Cartwright, 1971; Niedermeyer, Brown, & Sulzen, 1969). In fact, so few studies have shown a clear superiority for sequencing (Roe, 1962; Tobias, 1972) that the principle itself appears doubtful.

Commenting on the sequence research, Anderson (1967, 1970) suggested that positive effects of logical structure may be offset because practice on similar material is massed in most sequenced programs. Alternately, randomizing frames serves to space exposures to similar material, taking advantage of distributed practice effects. What happens is that disordered programs allow a larger time interval to occur between contact with frames having near identical content. Often, adjacent frames require very similar responses; consequently, it may be that distributing encounters with interrelated material leads to performance facilitation in programs in much the same manner as it does with larger units of text (e.g., Reynolds & Glaser, 1964). In many logically sequenced programs, however, practice density is high for like materials, and one might expect similar responses to interfere with one another at criterion recall. The argument then, is that any facilitation due to sequence may be counterbalanced by a massed practice decrement. Under these conditions, we would be correct in predicting a no-difference result when comparing sequenced and unsequenced programs. The authors can find no research that has tested this possibility.

Most studies manipulating sequence have varied only the order of frame presentation. Should distributed practice effects operate as we suppose, it would be difficult to investigate them within a single-program design, since all the subject matter is conceptually related. We reasoned that the separate in-

[1] This research was funded by Arizona State University Faculty Research Grant 7805-837-15-8 to the second author. A shorter version of this paper was read at the meeting of the American Educational Research Association in New Orleans, Louisiana, March 1973.

The authors wish to thank Mike Ash, Gary Estes, Bonnie Jensen, Don Milam, and Ron Zellner for their help in collecting data for this study. Our special thanks go to Alta May Miller for typing the manuscript.

[2] Requests for reprints should be sent to Raymond W. Kulhavy, College of Education, Arizona State University, Tempe, Arizona 85281.

fluence of distribution and sequence could be isolated by simply varying the location and order of the experimental frames throughout a body of nonrelated placebo text. Using this procedure, we could test differences due to both separation of frames containing like material and their order of presentation. Also, groups receiving only the experimental frames, in both logical and random order, would allow a direct test of effects due to program length and presence or absence of placebo material. Here, the distributed practice hypothesis predicts an interaction between distribution and order, with little difference between massed presentations but significant facilitation when the critical frames are distributed throughout the placebo text.

It is possible that massed presentation decreases criterion performance because of competition among contiguous responses. If recall is interfered with in this fashion, one would expect massed subjects to select a greater number of incorrect, similar responses on a multiple-choice test. On a constructed response measure, scores should be differentially higher for massed subjects, because interfering responses are not available for competitive recognition. We could test this assumption by including distractor choices that are selected from material adjacent to the correct response in the program for each multiple-choice item. Our prediction is that adjacent distractors will be chosen more often by the massed presentation learners. In the present study, the test mode variable was included to assess this hypothesis.

Finally, not only interference but the amount of attention the subject pays to the frames may be influenced by spaced encounters. A number of the nonsignificant studies have shown longer times for the disordered frame group. The argument for the distribution hypothesis would be buttressed if we could show that frame reading time is greatest for the distributed groups. It seems possible that the amount of time the subject spends reading may be inversely related to his exposure to like material prior to reading a particular frame.

Method

General Design

Two variables, practice distribution and frame order, were combined factorially to yield four experimental groups. A third variable, test mode, was varied within subjects across each factorial cell. The design was thus a 2 (Distributed vs. Massed Practice) × 2 (Logical vs. Random Order) × 2 (Multiple-Choice vs. Constructed Response Test) mixed analysis of variance with repeated measures on the test mode variable.

Three control groups completed the experimental design. The first of these (Control 1) studied only the experimental frames in normal order without seeing the placebo material. The second control (Control 2) was identical to the first, except that the frames were seen in a random order. The third group (Control 3) read only the placebo material.

Procedure

Subjects participated in groups of 8–58, with individuals from each condition in every session. Prior to handing out the materials, subjects were asked to define "myocardial infarction." Those learners who could do so successfully were dropped from the data analysis. Following this pretest, packets of material from the seven conditions were handed out randomly to participants. Control group packets were distributed at a slightly lower ratio than were the experimental packets. Each packet contained the appropriate program booklet prefaced by a sheet of instructions. Subjects were asked to read each program frame, respond to the questions, and record their time for completing each frame from a visible time board. Subjects were cautioned to read steadily through the program without returning to frames previously studied. No mention was made of the varying program sequences. Figures recorded from the time board were accurate to 10 seconds. When all subjects signified that they had read and understood the instructions, they were asked to begin.

As soon as he completed the program, the subject raised his hand and a monitor removed the text and gave him the constructed response form of the criterion test. The multiple-choice form was given to the subject as soon as he finished the constructed response version. No time limit was imposed on completion of either test form.

Subjects

The subjects were 176 undergraduates from Arizona State University. Twelve subjects were dropped from the study for failure to follow instructions or for prefamiliarity with the experimental subject matter. The final design contained 26 subjects in each of the four experimental groups and 20 in each of the controls.

Materials

The experimental program contained 48 frames, adapted from Tobias (1968), on the diagnosis of myocardial infarction from electrocardiograms. The experimental frames averaged 24 words in length and included line drawings as well as text. Each frame required the subject to respond either in writing or by constructing a line drawing. A validation study prior to the main experiment assured that (*a*) the program sequence was perceived as logical by the reader and (*b*) each of the 32 test questions could be answered correctly from reading the applicable program frames.

The placebo program consisted of three 20-frame sections on the respiratory system (Putnam, Lightfoot, & McDaniel, 1970), the muscles of the trunk, and the knee joint (McDaniel, Kindig, & Putnam, 1965). The 60 placebo frames averaged 27 words in length, contained the same number of illustrations as the infarction program, and were judged to have no content that overlapped with the experimental frames. The response requirements for the placebo program were essentially the same as those for the experimental program.

Within the four treatment groups, the experimental frames were either distributed evenly throughout the placebo material or massed together in one location. For the order conditions, experimental (E) frames were either random presentations or the logical presentations given in the original program sequence. Location of the massed frame presentations within the placebo (P) sections was counterbalanced (e.g., E–P_1–P_2–P_3; P_1–E–P_2–P_3; P_1–P_2–E–P_3; P_1–P_2–P_3–E) across each condition. In Control Groups 1 and 2, the experimental frames were given alone in either logical or random order. Control Group 3 read only the placebo material.

The two criterion tests consisted of 32 identical items. Questions in the constructed response version contained a blank for the subject to fill in; whereas the multiple-choice stems were followed by four response alternatives. About 15% of the criterion items required the subject to construct or identify an electrocardiogram drawing. In order to test the interference notion, each multiple-choice item contained a distractor selected from frames immediately adjacent to the one containing the correct answer in the original program. The remaining distractors were chosen from more distant positions.

Results

Table 1 presents criterion score means and standard deviations for both the experimental and control conditions. A 2 (Practice Distribution) × 2 (Frame Order) × 2 (Test Mode) analysis of variance on these data yielded significant main effects for order ($F = 3.99$, $df = 1/100$, $p < .05$) and test mode ($F = 34.41$, $df = 1/100$, $p < .01$). None of the remaining terms in this analysis reached significance.

To test for specific placebo material effects, a second unweighted-means analysis was performed on the combined experimental and control groups (Controls 1 and 2). This 2 (Placebo × No Placebo) × 2 (Frame Order) × 2 (Test Mode) analysis showed significant results only for the Frame Order main effect ($F = 5.66$, $df = 1/88$, $p < .05$). Apparently, whether or not the subject received the placebo frames had little effect other than to attenuate test mode differences. Differences between the ordered and randomized controls essentially replicated those of the experimental groups ($t = 2.57$, $df = 38$, $p < .05$). A 2 (Practice Distribution) × 2 (Frame Order) analysis of variance on the number of adjacent distractors chosen on the multiple-choice test yielded no significant results.

TABLE 1

Mean Criterion Scores by Sequencing and Test Mode Collapsed across Distribution

Sequencing of frame material	Experimental group		Control group			
			Experimental material		Placebo	
	Multiple choice	Constructed response	Multiple choice	Constructed response	Multiple choice	Constructed response
Logical						
M	19.53	17.08	20.81	18.75	1.25	.60
SD	6.64	6.15	7.20	6.41	3.10	.86
Random						
M	16.83	15.06	15.81	13.90		
SD	5.94	5.84	7.39	6.26		

TABLE 2

MEAN EXPERIMENTAL PROGRAM COMPLETION TIME IN MINUTES FOR SEQUENCE AND DISTRIBUTION CONDITIONS

Sequencing of frames	Material distribution		Control
	Massed	Distributed	
Logical			
M	23.33	22.20	26.77
SD	6.88	5.75	6.68
Random			
M	23.61	17.39	25.99
SD	7.42	5.04	8.13

Table 2 shows the means and standard deviations for time needed to complete only the experimental frames. A 2 (Practice Distribution) × 2 (Frame Order) analysis of variance for the experimental condition only gave significant results for the practice distribution main effect ($F = 8.42$, $df = 1/100$, $p < .01$) and the Practice Distribution × Frame Order interaction ($F = 4.02$, $df = 1/100$, $p < .05$). Combining both the experimental and controls (Controls 1 and 2) into a 2 (Frame Order × 2 (Placebo × No Placebo) unweighted-means analysis of variance yielded no significant effects.

A 2 (Practice Distribution) × 2 (Frame Order) analysis on errors made during the program yielded only one significant effect for the frame order variable ($F = 32.13$, $df = 1/100$, $p < .01$). Mean errors were 45.86 for ordered and 67.34 for random learners.

DISCUSSION

Clearly these results offer small support for the distribution hypothesis. Whether or not the experimental frames were massed together or distributed across the placebo program had little effect on amount of recall. Interestingly, criterion performance was significantly suppressed by the disordering procedure, regardless of whether or not the subject also read the placebo material. These results are, of course, exactly what one would predict if sequencing were of major instructional importance. The fact that interfering multiple-choice alternatives were selected no more often than other alternatives provides no power for the response competition argument.

An adequate explanation of these data seems to pivot on attending behavior. The amount of time spent reading experimental frames appears proportionately related to the criterion scores. In general, the massed presentation groups took longer to study the infarction material than did their distributed presentation counterparts. However, note the form of the Practice Distribution × Frame Order interaction. The differential between the pooled groups is almost entirely due to the rapid study time of subjects receiving the distributed random presentation. The criterion scores and error data follow this same trend.

From these results we could reason that when material is disordered *and* separated, it is more likely to attenuate the subject's inspection behavior. He is less willing to maintain high levels of attention when subject contiguity is disturbed along both these dimensions. Such disruption fails to occur when only disorder *or* distribution is manipulated. Our attention explanation generalizes to results of previous studies. Many programs used in this research were highly redundant in structure. In a markedly repetitive program, it would be difficult to distribute similar materials, simply because much of the total sequence contains the same information. Here we could expect minimal effects for disordering frames if the two-dimensional notion is valid. In a nonredundant program in which most of the material is critical for criterion performance, randomizing frames should have a measurable influence on recall, because randomizing in this case serves to distribute as well as disorder like material.

Tobias (1972) points out that sequence effects are most powerful when subjects have low preinstruction familiarity with difficult program content. Our two-dimensional hypothesis would predict exactly this result. When learners study a subject with which they have little prior experience, *all* of the material becomes critical for criterion performance. Disordering of frames should lead to strong sequence effects under these conditions because both disruption and distribution occur simultaneously, leading to the

same attenuation of inspection shown by our data.

If our argument is valid, the conclusions regarding instructional design become obvious. Sequence should receive close attention when preinstruction familiarity is low and the material is nonredundant. Such measures are obtainable for a given instruction package by pretesting and by some measure of "critical content" such as the blackout ratio (Kemp & Holland, 1966). Under conditions in which either, but not both, of these measures is high, disruption of "normal" order should have minimal effects.

REFERENCES

ANDERSON, R. C. Educational psychology. *Annual Review of Psychology*, 1967, **18**, 156–157.

ANDERSON, R. C. The control of student mediating processes during verbal learning and instruction. *Review of Educational Research*, 1970, **40**, 349–370.

CARTWRIGHT, G. P. The relationship between sequence of instruction and mental abilities of retarded children. *American Educational Research Journal*, 1971, **8**, 143–150.

GAGNÉ, R. M. *The conditions of learning*. New York: Holt, 1965.

KEMP, F. D., & HOLLAND, J. G. Blackout ratio and overt responses in programmed instruction: Resolution of disparate results. *Journal of Educational Psychology*, 1966, **57**, 109–114.

MCDANIEL, L. V., KINDIG, B. A., & PUTNAM, S. L. *Bones, joints and muscles of the human body*. New York: Macmillan, 1965.

NIEDERMEYER, F. C. The relevance of frame sequence in programmed instruction: An addition to the dialogue. *AV Communication Review*, 1968, **16**, 301–317.

NIEDERMEYER, F., BROWN, J., & SULZEN, B. Learning and varying sequences of ninth grade mathematics materials. *Journal of Experimental Education*, 1969, **37**, 61–67.

PUTNAM, S. L., LIGHTFOOT, J. D., & MCDANIEL, L. V. *Major systems of the human body*. New York: Macmillan, 1970.

REYNOLDS, J. H., & GLASER, R. Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology*, 1964, **55**, 297–308.

ROE, A. A. A comparison of branching methods for programming learning. *Journal of Educational Research*, 1962, **55**, 407–416.

TOBIAS, S. *The effect of creativity, response mode, and subject matter familiarity on achievement from programmed instruction*. New York: MSS Educational, 1968.

TOBIAS, S. The effect of sequence and familiarity with subject matter on achievement from programmed instruction. Paper presented at the meeting of the American Educational Research Association, Chicago, April 1972.

(Received January 3, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 62-66

EFFECTS OF TEST SCORE FEEDBACK ON IMMEDIATELY SUBSEQUENT TEST PERFORMANCE[1]

BRENT BRIDGEMAN[2]

University of Virginia

This experiment investigated the effects of success feedback, failure feedback, and no feedback on a scholastic aptitude test administered immediately after feedback was given. Sham scores, obstensibly from a similar test administered two days previously but actually randomly determined, were given to 233 male and female seventh-grade students immediately before taking a test that consisted of items from the nonverbal battery of the Lorge-Thorndike Intelligence Test. Subjects given success feedback (i.e., told that they had received a high score on the previous test) scored significantly higher ($p < .02$) than subjects given failure feedback. The group receiving no feedback was not significantly different from the average of the two feedback groups.

Test scores are widely assumed to be influenced by the motivational level and emotional state of the person taking the test. It is also clear that reporting test scores to students may affect their motivation, self-confidence, and anxiety level (Kirkland, 1971). Yet, there is little evidence directly relating the effects of reporting success or failure on one test to performance on an immediately subsequent academic test. A number of studies (Flook & Saggar, 1968; Hurlock, 1925; Means & Means, 1971; Sax & Reade, 1964) suggest that feedback from testing situations influences later test performance. In these studies, however, there is a period of time, ranging from a few days to a number of months, between the time of the feedback and the time of the later testing. The results, then, are due not only to possible differences in the affective state of the subject while taking the test but are presumably also strongly influenced by differential study habits in the interval between the report of the feedback and the subsequent testing. It is also of great interest, however, to determine the effects of feedback on test performance when study habits are not a potential intervening variable. For example, a test administrator would need to know whether giving feedback midway through a testing battery would influence scores on later tests in the battery or whether allowing students to grade one section of a test before proceeding to another section could influence their performance on subsequent sections.

The motivational effect of knowledge of results when there is no opportunity for study between the feedback and subsequent testing has been well documented (Locke, Cartledge, & Koeppel, 1968), but the tasks utilized in these studies are not comparable to the complex demands of an academic test situation. As Means and Means (1971) noted, most of the research in this area is limited to simple computational, physical, and verbal tasks that are not academic in nature. This distinction is important not only because of the different cognitive demands of these tasks but also because whether the subject perceives the same task as a research project or an educationally important test can affect his score (Katz & Greenbaum, 1963). The few studies that appear to use ed-

[1] This article is based on research submitted in partial fulfillment of the doctoral degree at the University of Wisconsin—Madison.

The author wishes to thank John Giebink, T. Anne Cleary, and Joel Levin for their assistance. Thanks also to the teachers and administrators at Wisconsin Dells and Baraboo Junior High Schools.

[2] Requests for reprints should be sent to Brent Bridgeman, Department of Educational Foundations, University of Virginia, Charlottesville, Virginia 22903.

ucational tests are not really relevant to the majority of academic tests, which are primarily power (not speed) oriented. Thus, for example, Osler (1954) claimed to assess feedback effects on a test of "intellectual performance." A group of students given sham feedback stating that their scores on a previous test in arithmetic put them in the lowest 10% of all students tested scored significantly lower on a subsequent test than another group that was told that they did very well. The experimental task, however, consisted of long-division problems administered to eleventh-grade students under a strict time limit. There is a strong implication that the experimental manipulation influenced only speed of performance. Similarly, Anderson, White, and Wash (1966) demonstrated the advantageous effects of success feedback over failure feedback on a 120-item multiple-choice test of addition and substraction of single-digit numbers with all subtraction differences being positive; obviously this was a nearly pure speed test with their sample of female college students.

The current experiment was designed to assess the immediate effects of success feedback (a high score on a previous test), failure feedback, or no feedback on a power test of academic aptitude administered to seventh-grade students under realistic conditions. Possible sex differences and differential effects of feedback on subjects of different ability levels were also investigated. Specifically, the following predictions were made:

1. Subjects given success feedback score higher than subjects given failure feedback. This is the contrast of primary interest and is clearly directional.

2. Subjects given no feedback differ in score from the average score of subjects given positive and negative feedback. This is a nondirectional test of feedback (regardless of type) versus no feedback.

3. Within males, the difference between the success and failure conditions is unequal in the high- and low-ability groups. This is a test of a specific feedback by ability interaction.

4. Within females, the difference between the success and failure conditions is unequal in the high- and low-ability groups. This is a test of a specific feedback by ability interaction.

5. The difference between the success and failure conditions will be unequal in males and females. This is a test of a specific feedback by sex interaction.

6. There is an overall sex difference.

7. Within males, high-ability subjects score higher than low-ability subjects. This is simply a check on the legitimacy of using the short pretest as a blocking variable.

8. Within females, high-ability subjects score higher than low-ability subjects. This is simply a check on the legitimacy of using the short pretest as a blocking variable.

Method

Subjects

The subjects were all of the seventh-grade students present on both testing days from two schools in small cities in central Wisconsin (N = 233). It was decided in advance that approximately this number of students was necessary in order to be able to detect differences between the positive and negative feedback conditions (the contrast of primary interest) of one-half of a standard deviation with a power of .80 when alpha was set at .05.

Materials

All test items on both the pretest and the criterion test were taken from the Lorge-Thorndike Intelligence Tests, Multi-Level Edition, Form 1, nonverbal battery.[3] This test was chosen because it is a widely used, well-constructed group intelligence test (Cronbach, 1960). The nonverbal battery was selected because it would appear to the student to be independent of vocabulary and reading ability, and hence students who were weak in those areas might still believe they could get high scores.

For the pretest, items were selected from the first subtest in the nonverbal battery. These items consist of three drawings with something in common, and the student must select from five choices which figure "goes with the first group."

Items for the criterion test were selected from Subtest 2 of the nonverbal battery and included all items from Levels D, E, and F (i.e., Items 15–60), which is the difficulty level appropriate for seventh-grade students. In these items, the subject is asked to figure out the way in which a series of

[3] Items used by permission of the publisher, Houghton Mifflin Company.

items is arranged and then to pick the next number from five choices. On both tests, the answers were recorded on machine-scorable answer sheets.

Procedure

The testing was done near the end of the school year when the students normally expect to be taking standardized tests. The students were notified ahead of time that they were to be tested on two days during the week, although the nature of the testing was not revealed. On the day of the initial testing, the experimenter told the children to read the instructions written on the front page of the test as he read them out loud. The instructions were as follows:

> This is a test of your ability to solve abstract problems. Since your school work often requires memorization rather than this type of thinking, your score on the test may not be directly related to how well you have done in school so far.

Standard directions (slightly modified, since different test booklets and answer sheets were used) for taking the test then followed. The time limit was 13 minutes, which is slightly more time than is required in the test manual. This was done to allow even the slow students time to at least attempt nearly all of the items, which was necessary if they were to believe that they had done very well on the test. The results of the test were used to divide the subjects of each sex into high-, medium-, and low-ability groups, with as nearly equal numbers as possible in each ability level (i.e., male and female groups were separately divided into thirds, based on the scores of the first test).

Two days after the first test, the second test was administered. The students were cautioned not to talk after receiving their test papers, and to leave the test booklets unopened on their desks. The top sheet contained only the student's name. After all of the papers had been passed out, the students were instructed to open their test booklets to the second page, where they found their "score" on the previous test and an interpretation of it. Actually, the scores were randomly assigned (in as nearly equal numbers as possible) within each sex and ability group. Of all the children tested, 77 received a score of 96; 77 received a score of 53; and 79 had a blank space after "score" on the interpretation sheet. On the interpretation sheet, 90 and above corresponded to the comment "Excellent! Your problem solving ability is among the best of all seventh-grade students." The score 59 and below corresponded to the comment "Poor. Your problem solving ability is among the worst of all seventh-grade students." The no-feedback group was told, as the tests were being distributed, that some of the papers had not yet been graded. The experimenter than instructed the subjects to read the instructions on the next page of the test booklet as he read them aloud. The instructions were as follows: "This is another test of your problem-solving ability. Like the other test you took, it is not related to how well you have done in school so far." The standard instructions (slightly modified) then followed. The time limit was 17 minutes, which is slightly more than recommended in the test manual, in order to minimize the effects of speed and further assure that it was basically a power test.

Design

The experimental design was a randomized block factorial with ability nested in sex. Sex was a nested factor, since ability grouping was done separately for males and females, and hence the cutoff scores for the ability levels were not identical in each sex.

Each of the eight planned comparisons was tested as a one degree of freedom contrast. The first hypothesis was of greatest interest and was tested with alpha set at .05. The other seven hypotheses were each tested with alpha set at .01 in an effort to keep the experimental error rate small.

Results

Scores consisted of the number of correct responses on the second test. Numbers, means, and standard deviations per cell for the test score data are reported in Table 1.

For each of the eight planned comparisons, the within cell mean square was 20.53 ($df = 1/215$). The major hypothesis that subjects given success feedback will score higher than subjects given failure feedback was supported ($F = 5.38$, $p < .012$, one-tailed). The no-feedback condition was neither significantly better nor worse than the average of the feedback conditions ($F < 1$). For both males and females the specified feedback by ability interactions were not significant ($F = 1.23$ and $F < 1$, respectively). The specified sex by feedback interaction was also not significant ($F < 1$). Females, on the average, scored higher than males ($F = 11.07$, $p < .001$). In both sexes, students selected as high scorers on the pretest scored significantly higher on the posttest than students selected as low scorers (for males, $F = 55.24$, $p < .001$, one-tailed, for females, $F = 30.93$, $p < .001$, one-tailed), indicating that the pretest scores served as a legitimate blocking variable.

Discussion

The results suggest that under realistic academic aptitude testing conditions, the

TABLE 1

NUMBERS, MEANS, AND STANDARD DEVIATIONS ON THE CRITERION TEST

Feedback condition	Males			Females			Total
	High ability	Middle ability	Low ability	High ability	Middle ability	Low ability	
Success							
n	13	13	13	12	12	14	77
M	22.62	18.46	16.08	23.08	22.25	18.00	19.99
SD	4.29	3.26	4.97	3.85	3.17	5.33	
Failure							
n	13	13	12	11	14	14	77
M	21.69	19.38	12.33	21.55	20.14	15.21	18.36
SD	5.41	4.44	6.38	3.24	3.66	4.21	
None							
n	13	14	11	14	14	13	79
M	21.85	16.00	14.36	23.00	20.14	17.23	18.91
SD	4.24	5.10	5.99	3.06	3.68	5.56	

student's perception of his prior success or failure influences his subsequent performance. The implications for the teacher or guidance counselor are clear; comparative feedback should be withheld until all tests in a battery have been administered, thus assuring that feedback from the first tests will not differentially affect scores on the later tests.

While there has been considerable interest in educational expectancy effects since the widely reported Rosenthal and Jacobson (1968) study, attention has focused almost exclusively on teacher expectancy. Perhaps because of the rather indirect linkage between the creation of expectancies in teachers and the ultimate test performance of their students, a number of attempts to replicate the general findings of Rosenthal and Jacobson have failed (Finn, 1972). The current experiment suggests that the more direct manipulation of self-expectancy, as through the report of the test score directly to the student, deserves considerably more attention from researchers interested in educational expectancy effects. Similarly, "halo" effects, in which initial perceptions of success or failure influence scoring of subsequent test responses, should be viewed as affecting not only test administrators but test takers as well. Thus, the person who fails initially may not only be scored lower because of his failure, but may actually perform more poorly than he could have because of his perception of himself as a failure.

Since there was no significant difference between the no-feedback group and the average of the two feedback groups, any definitive statement is unwise. The apparent location of the no-feedback group between the success and failure feedback groups at least suggests that knowledge of results of the "motivational" (i.e., not corrective) type discussed by Locke, Cartledge, and Koeppel (1968) should not be automatically assumed to be facilitating, but rather careful attention should also be given to whether the knowledge of results is likely to be perceived by the subject as indicative of success or failure.

The failure to find significant sex by feedback or ability by feedback interactions may be partially attributable to the available sample size, which was sufficient to provide the desired power only for the main effects; hence the power for the interaction hypotheses was less than adequate. Nevertheless, the data suggest some areas for further investigation. A possible ability by feedback interaction, especially in males, should be studied. For males in the high-ability group, the difference between success and failure feedback was only .93; while in low-ability males, it was 3.75, yielding a difference of 2.82. In females the same pat-

tern occurred, although the difference of the differences was only 1.26. It may be that bright students are more likely to try their best under any circumstances, while slower students, especially males, give up more easily unless they are working on tasks in which a report of past success makes them believe they have a chance to succeed. This speculation should be confirmed by further research, since it is only suggested, and not really supported, by the current investigation.

REFERENCES

ANDERSON, H. E., WHITE, W. F., & WASH, J. A. Generalized effects of praise and reproof. *Journal of Educational Psychology*, 1966, **57**, 169–173.

CRONBACH, L. J. *Essentials of psychological testing*. (2nd ed.) New York: Harper & Row, 1960.

FINN, J. Expectations and the educational environment. *Review of Educational Research*, 1972, **42**, 387–410.

FLOOK, A. J. M., & SAGGAR, V. Academic performance with, and without, knowledge of scores on tests of intelligence, aptitude, and personality. *Journal of Educational Psychology*, 1968, **59**, 395–401.

HURLOCK, E. B. An evaluation of certain incentives used in school work. *Journal of Educational Psychology*, 1925, **16**, 145–159.

KATZ, I., & GREENBAUM, C. Effects of anxiety, threat, and racial environment on task performance of Negro college students. *Journal of Abnormal and Social Psychology*, 1963, **66**, 562–567.

KIRKLAND, M. C. The effects of tests on students and schools. *Review of Educational Research*, 1971, **41**, 303–350.

LOCKE, E. A., CARTLEDGE, N., & KOEPPEL, J. Motivational effects of knowledge of results: A goal setting phenomenon? *Psychological Bulletin*, 1968, **70**, 474–485.

MEANS, R. S., & MEANS, G. H. Achievement as a function of the presence of prior information concerning aptitude. *Journal of Educational Psychology*, 1971, **62**, 185–187.

OSLER, S. F. Intellectual performance as a function of two types of psychological stress. *Journal of Experimental Psychology*, 1954, **47**, 115–121.

ROSENTHAL, R., & JACOBSON, L. *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston, 1968.

SAX, G., & READE, M. Achievement as a function of test difficulty level. *American Educational Research Journal*, 1964, **1**, 22–25.

(Received January 15, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 67-82

EFFECTS OF MODELS OF CREATIVE PERFORMANCE ON ABILITY TO FORMULATE HYPOTHESES[1]

NORMAN FREDERIKSEN[2] AND FRANKLIN R. EVANS

Educational Testing Service, Princeton, New Jersey

The effects of training procedures, ideational fluency, verbal ability, test anxiety, and sex on Formulating Hypotheses test performance were studied. Training consisted of presentation of models of "acceptable" responses that stressed either quantity or quality of performance. Both the quantity and quality models were found to be effective in modifying behavior in the expected direction. Ideational fluency was related to the number of hypotheses written, and verbal ability was related to scores reflecting quality of responses. Females were in general superior to males with respect to scores reflecting number of responses. Test anxiety was not significantly associated with performance. The effect of the training procedures was interpreted as changing standards with regard to what is "good enough" to report rather than changing ability.

A scientist who is trying to interpret the graphs or tables summarizing a research investigation may begin by making sure that he understands the study—what the variables mean, how the data were collected, the nature of any experimental treatments, etc. Then he may engage in speculations as to what may account for the results. He may begin by being skeptical of any apparently obvious interpretations and look for possible confounding, hidden sources of error, poor research design, or inappropriate methods of analysis. Or he may begin by indulging in unrestrained speculation about new theoretical implications and insights that may be suggested by the data. Eventually he will begin to operate under the constraint that whatever interpretation he seriously entertains should be consistent with all the data and with other information available to him. After engaging for a while in the process of generating explanatory concepts, on the one hand, and evaluating them against such criteria as their theoretical value and logical consistency, on the other, the scientist may be left with a fairly large number of hypotheses in mind that vary considerably with regard to probability of being correct. If at this point the scientist is asked to suggest interpretations of the finding, he may, depending upon his personal predilections, propose a large number of hypotheses, some of which are blue-sky speculation, or he may propose only a few ideas that meet rigorous subjective standards.

Formulating Hypotheses (Frederiksen, 1959) is a test that was developed in an attempt to measure abilities of the sort required of a research scholar who is trying to make sense out of research findings. Each item of the test consists of a graph or a table showing findings from an actual research study, and the task is to propose hypotheses to account for the findings. The sample item is a graph showing yearly rates of death from infectious diseases and from diseases of old age. The finding, printed below the graph, is as follows: "Rate of death from infectious diseases has decreased markedly since 1900, while rate of death from diseases of old age has increased." The

[1] This research was supported by the National Institute for Child Health and Human Development, Research Grant 5 PO1 HDO1762.

The authors wish to thank Albert E. Beaton and Charles E. Hall for their help with data analysis and Donald B. Rubin and William C. Ward for their critical review of the manuscript.

[2] Requests for reprints should be sent to Norman Frederiksen, Division of Psychological Studies, Educational Testing Service, Princeton, New Jersey 08540.

subject is provided with an answer sheet and instructed to write "short statements of hypotheses (possible explanations) which you think might account for, or help to account for, the finding." Sample answers are provided, including the following: "As treatment of infectious diseases has become more successful, more people survive to old age and, since everyone must die, death from diseases of old age increases"; "Compulsory inoculation laws have produced a continuous decline in incidence of infectious diseases and hence a decline in mortality"; and "More restrictive definitions of *infectious* have reduced the number of diseases so defined, while more diseases have been defined as 'old age' diseases." Items in the test itself present data showing, for example, that time lost from work stoppages was greatest in the summer months, that grades earned in college were on the average higher for younger than for older freshman students, and that during a certain postwar period employment decreased in mining and increased in all the other industries included in a survey. Thus the test attempts to simulate an aspect of the creative work of a scientist.

Applying our armchair analysis of a scientist's thinking to Formulating Hypotheses items suggests that several cognitive abilities may be involved in taking the test. One kind of ability would presumably be ideational fluency, the ability to generate many ideas, including some unusual or original ideas. But because of the constraint that the solutions must be potentially useful in accounting for the finding, other cognitive abilities such as reasoning, memory, and verbal ability would be required. This implies that a proper balance must be maintained between generating many ideas (including some that may be inconsequential) and discarding ideas (some of which may be valuable) that fail to meet some standard of rigor or probability of veridicality. An attempt to account for individual differences in arriving at such a balance gets us into the domain of personality. Those who are willing to propose interesting though implausible hypotheses may be people with a high degree of self-confidence or a lack of concern about opinions of others. Those who discard all but the most obvious hypotheses may be anxious or defensive. Or another possible way to account for the contrast is in terms of the kind or level of standards of excellence one has learned to set for himself—standards that determine for each individual when he says, "That's good enough." This study represents an attempt to investigate the role of some of these factors in the process of formulating hypotheses, using Formulating Hypotheses as a source of the dependent variables.

There has been a good deal of controversy regarding the correlation between creativity and intelligence, the usual finding being that the correlation is positive but low. But the measures of "creativity" that are typically employed in these studies are scores on tests that are more properly labeled fluency tests, reflecting number of responses and number of unusual responses. Measures of quality have rarely been used, perhaps because the nature of the tasks involved does not elicit performance of sufficient breadth and variety to make possible reliable judgments of quality. Formulating Hypotheses items do elicit responses that clearly vary in quality as well as quantity, and one might expect scores based on quality judgments to correlate with the convergent thinking tasks involved in the typical intelligence test, while number of responses might correlate with the measures of fluency. One of the purposes of the present study was to develop quality as well as quantity scores for Formulating Hypotheses and to investigate their relationships to verbal ability as well as to tests of fluency.

In an earlier study (Klein, Frederiksen, & Evans, 1969), it was shown that the number of responses to Formulating Hypotheses could be increased merely by presenting to subjects models in the form of sets of responses containing substantially more hypotheses than subjects typically write. It was suggested that the effect of the models was probably not to change ability but to alter subjects' standards with regard to what constitutes satisfactory performance. Such an interpretation is consistent with the findings of Levy (1968), who showed that present-

ing samples of test responses as a way of defining a role that subjects were asked to adopt was effective in increasing the number of "original" (uncommon) responses to a word association task. The finding is also consistent with a study by Stratton and Brown (1972); they used a "judgment-training" technique that involved presenting examples of good and bad responses (as well as practice with criteria for evaluation of responses) and found that the training substantially increased the quality of the responses (plot titles). Stratton and Brown appear to attribute the improvement to an increase in ability to evaluate, but it seems equally likely that the improvement can be attributed to altered standards. The study to be reported here extends this line of research by using both quantity and quality models in an investigation of their main effects and their interactions with other factors.

The study by Klein et al. was primarily concerned with the role of anxiety in learning to formulate hypotheses. The specific hypothesis investigated was that the performance of anxious students would improve more than that of less anxious students as a result of the training with quantity models. The idea was that performance on such a task involves some self-censorship of ideas, especially by anxious people; that is, people are likely to have more ideas than they report because they do not write down ideas that they think aren't "good enough." The presentation of quantity models of "acceptable" responses, it was thought, would decrease the amount of censorship, thereby increasing the number of responses, and this effect would be greater for anxious than for less anxious individuals. The number of responses did increase as a result of the treatment, but since the interaction of treatment and anxiety was not significant, the hypothesis that the treatment would be more effective for anxious subjects was not confirmed.

An unexpected finding was that the relationship of anxiety to the number of hypotheses produced was curvilinear; poorest performance was associated with a middle level of anxiety and best performance, with a low level of anxiety. This result is not predicted by drive theory and is contrary to Stennett's (1957) suggestion that the relationship between performance and drive level is nonmonotonic with the shape of an inverted U.

The purposes of the present study, then, were (*a*) to investigate the relationships of both quantity and quality scores on Formulating Hypotheses to measures of fluency and verbal ability, (*b*) to see if "quality" models would be successful in improving quality of performance, and (*c*) to attempt to replicate two earlier findings—the effects of "quantity" models on number of responses and the curvilinear relationship of anxiety to performance.

Method

Subjects

The subjects in this study were 395 paid volunteers, mostly freshmen, at two state colleges in Pennsylvania. Approximately half were males and half females.

Measures

Formulating Hypotheses. In the previous study by Klein et al. (1969), two scores based on Formulating Hypotheses were used as dependent variables: number of hypotheses and number of acceptable hypotheses, and they were also used in the present investigation. Number of responses (Formulating Hypotheses Score 1) is the total number of nonduplicate hypotheses written by a subject, and number of acceptable responses (Formulating Hypotheses Score 2) is a subset of Formulating Hypotheses Score 1 responses that includes only ideas previously judged by a panel to be "acceptable." This score was used, in spite of its experimental dependence on Formulating Hypotheses Score 1, in order to make possible a more exact replication of certain aspects of the previous study.

Formulating Hypotheses Score 3 is a new score called average judged quality of the responses. It is the average of ratings made by two scorers (using a 9-point scale) of the quality of the responses, quality being defined in a sense that was consistent with the instructions to the examinees who took the Formulating Hypotheses test.

Formulating Hypotheses Score 4 is called average scale value. It is another quality score, obtained by a method that makes it relatively independent of such qualities as length, handwriting, or grammatical correctness. The method made use of a master list of the hypotheses written by students, as derived from a content analysis. A panel of judges made evaluations of the hypotheses on

this list, and scale values were assigned on the basis of these ratings. The scorer's task was merely to decide which listed hypothesis, if any, was similar to the one being scored, and to record the number of that listed hypothesis. The appropriate scale value was assigned to each response by computer.

Formulating Hypotheses Score 5 is average number of words per response. It was included to provide some insight into the processes involved in modifying behavior by presenting models and the cues used by scorers in evaluating responses.

An attempt to develop another score that would represent the rarity, unusualness, or originality of subjects' responses was unsuccessful because of low reliability.

Each of five Formulating Hypotheses scores was obtained separately for each of the seven Formulating Hypotheses items; this made it possible to obtain scores for subsets of items. Two subsets were used: (*a*) the first two items (which were completed before any models were presented) and (*b*) the last five items (which were all susceptible to influences of the experimental treatments). The five Formulating Hypotheses scores based on Items 1 and 2 are called Formulating Hypotheses pretest scores and were used as covariates in the statistical analysis. The five Formulating Hypotheses scores based on Items 3–7 were used as dependent variables. Unit weights were used in obtaining the pretest scores. The weights used in obtaining the five-item composite scores were their loadings on the first principal component resulting from a principal-axes factor analysis of the intercorrelations of the five items. Five such factor analyses were done, one for each of the five Formulating Hypotheses scores.

Consequences test. Consequences is one of the tests used by Guilford (1967) to measure divergent production. Each item presents a hypothetical situation (e.g., "What would be the results if people no longer needed or wanted sleep?"), and the task is to list possible consequences of that situation. Two scores were obtained, using Guilford's scoring method: Consequences-Obvious and Consequences-Remote, representing Divergent Production of Semantic Units and Divergent Production of Semantic Transformations, respectively, in the structure-of-intellect model. These scores were used to measure transfer of experimental effects to another task involving divergent production and were treated as two more dependent variables.

Cognitive tests. Two tests from the *Kit of Reference Tests for Cognitive Factors* (French, Ekstrom, & Price, 1963) were also used: Advanced Vocabulary, a 36-item multiple-choice synonyms test, and the Theme test, which requires the subject to write two themes, the score being merely the number of words written. These two tests measure the factors of verbal comprehension and ideational fluency, respectively, according to the French et al. manual. In the structure-of-intellect model, the Theme test represents Divergent Production of Semantic Units.

Test anxiety. The same inventory used in the previous study provided the test anxiety score; this scale contains items from Harleston's (1962) measure of test anxiety and the items from the Alpert-Haber (1960) debilitating anxiety scale.

Procedure

The procedure used was very similar to that employed in the previous experiment, except that there were two experimental treatments instead of one, and the data were obtained in one long evening session rather than in three separate sessions.

The experimental treatments consisted of providing models of acceptable performance at the completion of each Formulating Hypotheses item (except the first) in the form of a list of hypotheses pertaining to that item that were supposedly written by college students. One treatment (the quantity treatment) was essentially the same as that used in the first study; it consisted of providing a fairly long list (18–26) of "acceptable hypotheses" illustrating ideas that the subject might have written in response to the preceding item. The other treatment (quality) was similar except that each list of acceptable hypotheses included only the best ideas, carefully worded. The quality list typically included only six or seven hypotheses, and they were somewhat longer than those used for the quantity models.

Members of the control group received no models; instead, they were given various questionnaires to occupy their time in what appeared to be a relevant way.

The lists were intended to provide models of performance that subjects would try to emulate in one way or another. Both the quality and quantity materials are believed to show a rather striking contrast to the work of most students: The quantity list contained many more responses than the average subject wrote, while the quality lists were noticeably superior in quality, both with respect to ideas and wording. In order to insure that the lists were read, the subject was instructed to study the list carefully, then to go back to his own list to make any revisions or additions he felt would improve the list. (The Formulating Hypotheses answer sheet produced a copy of the subject's responses. The original was removed before the feedback materials were presented, and the revisions were made on the copy. Only the original was used in scoring.)

All subjects from a given college were seated together in a large room, and all three treatment groups were handled simultaneously. Assignment of treatments to subjects was accomplished by handing to every third subject as he entered the room an envelope containing materials for one particular treatment. All documents were numbered as shown in Table 1, and instructions were given by referring to document numbers. The document in use at a particular period was taken by the subject from the top of the pile in his envelope and was placed at the bottom of the pile

TABLE 1

SEQUENCE OF PRESENTATIONS

Document number	Group			Time (in minutes)
	Control	Quality	Quantity	
1		Personality Inventory		Untimed
2		Advanced Vocabulary Test		8
3		Theme Test		8
4		Formulating Hypotheses practice item		Untimed
5		Formulating Hypotheses Item 1		10
6		Formulating Hypotheses Item 2		10
7	Questionnaire	Quality models	Quantity models	7
8		Formulating Hypotheses Item 3		10
9	Questionnaire	Quality models	Quantity models	7
10		Formulating Hypotheses Item 4		10
11	Questionnaire	Quality models	Quantity models	7
12		Formulating Hypotheses Item 5		10
13	Questionnaire	Quality models	Quantity models	7
14		Formulating Hypotheses Item 6		10
15	Questionnaire	Quality models	Quantity models	7
16		Formulating Hypotheses Item 7		10
17		Consequences Test		20

in the envelope when completed. Subjects knew that the materials were not identical for everyone, but they did not know the nature or purpose of the different treatments.

Statistical Analysis

Means, standard deviations, and intercorrelations of all variables were computed for all subjects combined and separately for the three experimental groups. Reliabilities were computed where possible.

Because of the relatively large number of dependent measures, the method of analysis chosen was multivariate analysis of covariance using a computer program by Clyde, Cramer, and Sherrin (1966). There were seven dependent measures: the five Formulating Hypotheses scores based on Items 3–7, Consequences-Obvious and Consequences-Remote. Since there were small differences between the two colleges in student performance, one covariate was the dichotomy *college attended.* Other covariates were the five Formulating Hypotheses pretest scores. These five scores were appropriately used as control variables whenever we were interested in *change* in performance on Formulating Hypotheses (i.e., when investigating effects of treatments).

The design factors were treatment, sex, vocabulary, ideational fluency, and test anxiety. However, it was not possible to use all five design factors in one analysis because some cell frequencies became too small. Instead, three separate multivariate analyses of covariance were performed with overlapping factors. The three analyses employed the following design factors, with number of levels within each factor shown in parentheses: (*a*) Treatment × Sex × Ideational Fluency × Test Anxiety (3 × 2 × 2 × 3), (*b*) Treatment × Sex × Ideational Fluency × Vocabulary (3 × 2 × 3 × 2), and (*c*) Treatment × Sex × Vocabulary × Test Anxiety (3 × 2 × 2 × 3). Three levels of test anxiety and of ideational fluency were used in order to make possible the detection of nonlinear relationships. Each of the three designs was used once with college attended as the only covariate and once with the five Formulating Hypotheses pretest scores used as covariates in addition to college attended.

The multivariate analysis of variance model employed in the analysis first removes variance attributable to the four-way interaction, then lower order interactions, and finally it deals with main effects. Each main effect was computed so that it was orthogonal to all other main effects and all interactions. Thus the R^2 may be interpreted as the percentage of variance uniquely attributable to the factor under consideration.

RESULTS

Intercorrelations and Reliabilities

Table 2 presents the means, standard deviations, intercorrelations, and reliabilities of the variables used as covariates and as design variables, using data for all subjects combined. Since the treatments may affect intercorrelations involving dependent variables, the intercorrelations of dependent variables are shown separately in Table 3 for the three treatment groups as well as for the total group; Table 4 shows the correlations of the dependent variables with the

TABLE 2

INTERCORRELATIONS AND RELIABILITIES OF INDEPENDENT VARIABLES

Measure	1	2	3	4	5	6	7	8	9
					Correlation				
1. Sex[a]	—	.12	.17	.26	.13	.22	.20	.05	.15
2. Test anxiety		.87[b]	−.16	.06	−.00	−.03	−.04	−.03	−.03
3. Vocabulary			.68[c]	.10	.05	.11	.15	.04	.14
4. Ideational fluency				.78[c]	.34	.28	.06	.05	.15
5. FH Pretest Score 1					.54[c]	.79	−.06	.08	−.20
6. FH Pretest Score 2						.44[c]	.45	.28	.02
7. FH Pretest Score 3							.54[c]	.53	.38
8. FH Pretest Score 4								.52[c]	.22
9. FH Pretest Score 5									.72[c]
					M and *SD*				
M	1.5	34.4	14.9	151.3	15.9	11.7	9.3	11.9	39.1
SD	.5	16.0	4.7	35.2	5.0	4.5	2.2	2.6	12.4

Note. To be significantly different from zero, r must equal .10 ($p < .05$) or .13 ($p < .01$). Reliabilities are in the main diagonal. Abbreviation: FH = Formulating Hypotheses. $N = 395$.

[a] Males = 1; females = 2.

[b] Reliability is the average interitem correlation corrected for length.

[c] Reliability is the correlation of separately timed halves corrected for double length.

other variables for the three treatment groups and for the total group.

Reliabilities of the various measures are shown in the main diagonals of Tables 2 and 3. (Reliabilities are based on the total group.) The correlations between the two-item and the five-item tests are shown in Table 4; these may be thought of as alternate form reliabilities. These correlations are, of course, attenuated by the fact that one of the two tests contains only two items. Formulating Hypotheses Score 5 (average number of words) is the most reliable and Formulating Hypotheses Score 4 (average scaled value) the least reliable of the Formulating Hypotheses scores. While the two-item pretest is adequate for use as a control variable, a longer test would obviously have been better. The data indicate that highly reliable measures of Formulating Hypotheses performance can be built by using a sufficient number of items.

Intercorrelations of Formulating Hypotheses test scores (see Table 3) show that the Formulating Hypotheses Score 1 (number of responses) and Formulating Hypotheses Score 2 (number of acceptable responses) are highly correlated, as is to be expected since Formulating Hypotheses Score 2 is based on a subset of the responses contributing to Formulating Hypotheses Score 1. The two scores designed to measure quality—Formulating Hypotheses Scores 3 and 4—are also highly correlated, in comparison with their reliabilities, and Score 3 has a substantial correlation with Formulating Hypotheses Score 2 (number of acceptable hypotheses), which reflects quality as well as number of responses. The correlation between Formulating Hypotheses Score 3 (average rated quality) and Formulating Hypotheses Score 5 (average number of words) suggests either that raters tend to be impressed by long responses or that a more lengthy response is necessary for higher quality.

Since scores on the five-item Formulating Hypotheses test are influenced by the experimental treatments, it is important to look at their intercorrelations and correlations with other variables separately for the three treatment groups. Table 3 includes the intercorrelations of the dependent variables for the three treatment groups. There were some differences in correlations that might

TABLE 3

INTERCORRELATIONS OF DEPENDENT VARIABLES FOR TOTAL GROUP AND FOR THE THREE TREATMENT GROUPS

Measure	1	2	3	4	5	6	7
Correlation							
1. FH Score 1	.80[a]	.72	−.18	.01	−.23	.28	.34
		.73	−.13	−.05	−.21	.34	.38
		.67	−.16	.03	−.08	.30	.41
		.73	−.21	.04	−.30	.30	.29
2. FH Score 2		.67	.43	.31	−.07	.22	.19
			.47	.29	−.02	.26	.24
			.49	.40	.10	.23	.20
			.39	.27	−.22	.21	.15
3. FH Score 3			.60	.53	.33	−.06	−.14
				.57	.36	−.06	−.15
				.59	.40	−.04	−.20
				.41	.21	−.08	−.07
4. FH Score 4				.48	.14	−.04	−.05
					.18	−.06	−.11
					.15	−.02	−.01
					.09	−.01	−.02
5. FH Score 5					.87	−.09	−.06
						.10	−.08
						−.18	.02
						−.26	−.12
6. Consequences-Obvious						.85	.15
							.12
							.20
							.10
7. Consequences-Remote							.75
M* and *SD							
M	29.3	18.9	14.4	17.4	62.4	42.8	16.7
	29.4	18.7	14.0	17.3	62.5	43.3	17.0
	27.1	18.0	14.8	17.4	65.6	44.2	16.7
	31.4	19.9	14.3	17.5	59.3	40.9	16.5
SD	8.0	5.6	2.7	2.1	19.5	15.1	8.0
	7.9	5.9	2.8	2.2	22.0	14.0	7.6
	6.5	4.5	2.8	2.3	17.9	16.5	8.4
	8.8	6.0	2.5	1.9	17.8	14.7	8.0

Note. The first entry in each cell is the correlation for the total group. The next three entries are the correlations for control, quality, and quantity treatment groups, in that order. The *n*s for the four groups are, respectively, 395, 134, 129, and 132. For the total group, *r* must equal .10 ($p < .05$) or .13 ($p < .01$) to be significantly different from zero. For the treatment group, *r* must equal .17 ($p < .05$) or .22 ($p < .01$) to be significantly different from zero. Abbreviation: FH = Formulating Hypotheses.

[a] Reliabilities (shown in the diagonal) are the average item intercorrelation within the total group corrected for length by the Spearman-Brown formula.

TABLE 4

CORRELATIONS OF INDEPENDENT VARIABLES WITH DEPENDENT VARIABLES FOR TOTAL GROUP AND FOR THE THREE TREATMENT GROUPS

Independent variable	Dependent variable							*M*	*SD*
	FH Score 1	FH Score 2	FH Score 3	FH Score 4	FH Score 5	Consequences-Obvious	Consequences-Remote		
Sex	.12	.14	.06	.05	.12	.21	−.07	1.5	.5
	.00	.08	.11	.08	.20	.38	−.07	1.5	.5
	.19	.21	.08	−.05	.16	.15	−.07	1.5	.5
	.20	.17	−.01	.13	−.04	.12	−.08	1.5	.5
		−.08	−.05	.02	−.05				
Test anxiety	−.03					.05	−.07	34.4	16.0
	−.11	−.06	.09	.20	.01	.06	−.27	32.3	17.4
	.06	−.05	−.15	−.11	−.17	.10	.04	35.2	16.2
	−.01	−.16	−.14	−.09	−.01	.00	.06	35.7	14.0
		.22	.18	.13	.12				
Vocabulary	.17					.05	.13	14.9	4.7
	.16	.23	.14	.11	.13	.08	.19	14.9	4.7
	.25	.19	.14	.12	.20	.05	.16	14.9	5.0
	.13	.24	.29	.18	.02	.03	.03	15.0	4.5
Ideational fluency	.32	.23	−.05	−.08	.08	.43	.28	151.3	35.2
	.36	.28	.06	−.06	.15	.47	.29	153.5	35.6
	.32	.23	−.08	−.12	.06	.47	.30	148.1	36.9
	.29	.17	−.12	−.04	.02	.37	.24	152.1	32.9
FH Pretest Score 1	.51	.38	−.10	.03	−.15	.23	.30	15.9	5.0
	.62	.38	−.16	.02	−.13	.26	.35	16.2	5.1
	.43	.39	.01	.08	−.10	.21	.23	16.0	4.9
	.53	.41	−.14	−.01	−.24	.22	.32	15.4	5.1
FH Pretest Score 2	.42	.46	.16	.14	−.02	.18	.20	11.7	4.5
	.48	.41	.03	.12	−.03	.20	.20	11.9	4.3
	.39	.53	.30	.23	.07	.12	.16	11.8	4.6
	.45	.50	.15	.05	−.11	.23	.24	11.4	4.5
FH Pretest Score 3	.04	.30	.45	.23	.27	−.01	−.07	9.3	2.2
	.06	.29	.36	.25	.24	.08	−.07	9.3	1.9
	.09	.41	.54	.33	.35	−.13	−.02	9.3	2.3
	−.01	.24	.45	.11	.26	.05	−.12	9.4	2.4
FH Pretest Score 4	.00	.13	.26	.23	.16	.01	−.01	11.9	2.6
	−.07	.07	.21	.30	.19	.01	.03	12.2	2.4
	−.03	.22	.42	.33	.17	−.09	−.12	11.6	3.0
	.08	.13	.15	.00	.12	.14	.08	11.8	2.5
FH Pretest Score 5	−.17	−.01	.33	.13	.74	−.04	−.04	39.1	12.4
	−.15	.03	.36	.10	.80	.08	.01	39.1	12.2
	−.13	.07	.32	.18	.70	−.13	−.06	39.4	13.1
	−.22	−.11	.30	.10	.74	−.06	−.08	38.7	12.0

Note. The first entry in each cell is the correlation for the total group. The next three entries are the correlations for control, quality, and quantity treatment groups, in that order. The *n*s for the four groups are, respectively, 395, 134, 129, and 132. For the total group, *r* must equal .10 ($p < .05$) or .13 ($p < .01$) to be significantly different from zero. For the treatment group, *r* must equal .17 ($p < .05$) or .22 ($p < .01$) to be significantly different from zero. Abbreviation: FH = Formulating Hypotheses.

be attributable to treatments. For example, correlations involving Formulating Hypotheses Score 5 (number of words) were apparently reduced or made negative by the quantity feedback. However, the differences are not great, and the pattern is generally the same for all three groups. The correlations with independent variables (see Table 4) do not differ greatly for the three groups. It is therefore concluded that factor structure was not substantially altered by the treatments and multivariate analysis of covariance results would, from this point of view, be interpretable.

Canonical Variates

Table 5 shows the correlations of the dependent variables with the four significant canonical variates (out of six computed) that were obtained by using college attended *and* the five Formulating Hypotheses pretest scores as covariates and without including any of the design factors. The canonical correlations (shown in the row labeled R) are correlations between the best weighted combination of dependent variables and the best weighted combination of covariates. These canonical variates are orthogonal, and they result from a step-down model (variance attributable to the first canonical variate is removed before computing the second, etc.). The correlations of dependent variables with the canonical variate may be used like factor loadings to interpret the canonical variates.

The first canonical variate was obviously defined by Formulating Hypotheses Score 5 (average number of words). The canonical correlation between covariates and the dependent variables was high, about .75, no doubt in part because of the higher reliability of Formulating Hypotheses Score 5, which correlates about .98 with the canonical variate. The only other dependent variable with appreciable correlations was Formulating Hypotheses Score 3, the average quality rating. These correlations suggest that there may be a tendency for raters to give higher ratings to the longer responses, or possibly, longer responses are necessary for statements of high quality.

Canonical Variate 2 was defined mainly by Formulating Hypotheses Scores 1 and 2 (number of hypotheses and number of acceptable hypotheses). Sizable correlations also occurred for Consequences-Remote, but not for Consequences-Obvious, which suggests that the quantity of production on the Formulating Hypotheses test involves divergent production of semantic transformations more than semantic units.

Canonical Variate 3 was mainly correlated with Formulating Hypotheses Score 3, the average quality rating. Other substantial positive correlations were found for Formulating Hypotheses Score 2 (number of acceptable responses) and Formulating Hypotheses Score 4 (average scale value) reflecting the quality component in both these scores. In contrast to Canonical Variate 2, the correlations with Consequences-Remote were negative, suggesting that the Formulating Hypotheses quality scores were quite different from the quantity scores with respect to the influence of fluency.

The last canonical variate, which was barely significant, was correlated mainly with Formulating Hypotheses Score 4, the average scale value. Formulating Hypotheses Score 4 is the least reliable Formulating Hypotheses score, and a good deal of the variance attributable to it had already been allocated to Canonical Variate 3. Thus there appear to be three major components in the domain of the dependent variables: length of responses, number of responses, and quality of responses.

When college attended was used as the only covariate, the canonical variate was significant ($p < .01$), showing that college attended was significantly related to the dependent measures; but the canonical correlation was only about .25. The correlations with the canonical variate may be interpreted as showing that for some reason one college was superior with regard to the quality score (Formulating Hypotheses Score 3), and the other was superior on number of hypotheses (Formulating Hypotheses Score 1) and on Consequences-Remote.

Multivariate Analysis

Tables 6, 7, and 8 present the salient findings of six multivariate analyses of covariance. The analyses differed with respect to

TABLE 5

CORRELATIONS OF DEPENDENT VARIABLES WITH CANONICAL VARIATES

Dependent variable	Canonical Variate 1			Canonical Variate 2			Canonical Variate 3			Canonical Variate 4		
	T×S×I×A	T×S×I×V	T×S×V×A	T×S×I×A	T×S×I×V	T×S×V×A	T×S×I×A	T×S×I×V	T×S×V×A	T×S×I×A	T×S×I×V	T×S×V×A
						Correlation						
FH Score 1	−.21	−.29	−.24	**.91**	**.84**	**.90**	−.14	−.33	−.03	−.04	−.12	−.07
FH Score 2	−.03	−.09	−.08	**.83**	**.88**	**.76**	.46	.30	.55	−.16	.08	−.16
FH Score 3	.34	.34	.32	.13	.31	.02	**.88**	**.84**	**.88**	.17	.19	.20
FH Score 4	.12	.10	.10	.21	.23	.11	.34	.28	.32	**.77**	**.88**	**.79**
FH Score 5	**.98**	**.97**	**.98**	.03	.06	.08	−.19	−.22	−.16	−.02	.01	−.05
Consequences-Obvious	−.09	−.14	−.09	.12	.04	.28	−.05	−.10	−.16	.19	.23	.11
Consequences-Remote	−.07	−.08	−.08	.40	.35	.48	−.38	−.42	−.34	.13	.10	.15
						Multivariate significance level						
R	.76	.74	.75	.52	.50	.55	.42	.40	.40	.18	.20	.18
F	13.62	12.71	13.78	7.03	6.43	7.36	4.65	4.23	4.21	1.83	1.85	1.64
$p <$	.001	.001	.001	.001	.001	.001	.001	.001	.001	.039	.036	.075

Note. Covariates are college attended and five Formulating Hypotheses (FH) pretest scores. Abbreviations: T = Treatment; S = Sex; A = Anxiety; I = Ideational fluency; and V = Vocabulary. Highest correlations with each canonical variate are shown in boldface.

TABLE 6

RESULTS OF THE MULTIVARIATE ANALYSIS OF COVARIANCE FOR TREATMENT, SEX, IDEATIONAL FLUENCY, AND TEST ANXIETY

Effect	Multivariate significance level	*M*		Correlation with canonical variate		*p* for univariate tests
Treatment[a]	$F = 4.377$ $p < .001$ $R = .369$	Control group Quality treatment Quantity treatment	.03 −.47 .44	FH Score 1 FH Score 5 FH Score 2 Consequences-Obvious FH Score 3	.76 −.46 .46 −.31 −.23	.001 .003 .003 .065 .030
Sex[b]	$F = 3.637$ $p < .001$ $R = .260$	Male Female	−.27 .27	Consequences-Remote FH Score 2 Consequences-Obvious FH Score 5	−.49 .45 .44 .43	.013 .023 .024 .030
Ideational fluency[b]	$F = 13.654$ $p < .001$ $R = .462$	Low High	−.47 .47	Consequences-Obvious FH Score 1 Consequences-Remote FH Score 2	.82 .51 .48 .33	.001 .001 .001 .001
Test anxiety[b]	$F = 1.461$ $p < .120$ $R = .220$	Low Middle High	.26 .01 −.27	Consequences-Remote FH Score 2 FH Score 5 FH Score 1	.57 .50 .36 .31	.033 .104 .259 .337
Treatment × Test Anxiety interaction[a]	$F = 1.274$ $p < .155$ $R = .211$			FH Score 4	—	.014

[a] College attended and the five Formulating Hypotheses (FH) pretest scores used as covariates.
[b] College attended used as the only covariate.

the combination of design factors employed, as was described previously, and with respect to the covariates used (college attended only or college attended and the five Formulating Hypotheses pretest scores). Table 6 shows the results for the design factors treatment, sex, ideational fluency, and test anxiety; Table 7, the results for treatment, sex, ideational fluency, and vocabulary; and Table 8, the results for treatment, sex, vocabulary, and test anxiety. For treatment, and for interactions involving treatment, the results are reported only for the analyses in which Formulating Hypotheses pretest scores as well as college attended are used as covariates. For the remaining design factors and interactions, results are reported for analyses in which college attended is the only covariate. The Formulating Hypotheses pretest scores are used as covariates for treatment effects because we wish to use a measure of *change* in evaluating the effects of the quality and quantity models. Results for all the main effects are reported, but results for interactions are reported only if the significance level reaches $p < .05$ for either multivariate or univariate tests. The results shown for the main effects were computed in such a way that R^2 can be interpreted as the percentage of variance uniquely attributable to a particular factor.

A word about the contents of Tables 6, 7, and 8: The first column indicates what main effect (or interaction) is described in the corresponding row of the table. The second column shows the overall multivariate F ratio and its related p and R values. (The second canonical variate was not significant in any instance.) Except in the case of an interaction, mean scores on the first canonical variate for the appropriate subgroups are shown in the next column. (The grand mean is set at zero.) The next column shows the salient correlations of dependent variables with the first canoni-

TABLE 7

RESULTS OF THE MULTIVARIATE ANALYSIS OF COVARIANCE FOR TREATMENT, SEX, IDEATIONAL FLUENCY, AND VOCABULARY

Effect	Multivariate significance level	*M*		Correlation with canonical variate		*p* (for univariate tests)
Treatment[a]	$F = 4.455$	Control group	.04	FH Score 1	.74	.001
	$p < .001$	Quality treatment	−.46	FH Score 2	.45	.003
	$R = .371$	Quantity treatment	.42	FH Score 5	−.44	.004
				Consequences-Obvious	−.32	.061
				FH Score 3	−.22	.029
Sex[b]	$F = 3.286$	Male	−.24	Consequences-Remote	−.54	.009
	$p < .002$	Female	.24	Consequences-Obvious	.53	.011
	$R = .248$			FH Score 2	.35	.094
				FH Score 5	.33	.115
Ideational fluency[b]	$F = 8.459$	Low	−.65	Consequences-Obvious	.81	.001
	$p < .001$	Middle	−.05	FH Score 1	.51	.001
	$R = .497$	High	.70	Consequences-Remote	.46	.001
				FH Score 2	.30	.005
Vocabulary[b]	$F = 4.827$	Low	−.28	FH Score 3	.80	.001
	$p < .001$	High	.28	FH Score 4	.66	.001
	$R = .296$			FH Score 2	.58	.001
				FH Score 5	.51	.003
Sex × Vocabulary interaction[b]	$F = 1.598$			FH Score 2	.64	.032
	$p < .135$			FH Score 4	.61	.041
	$R = .176$			FH Score 3	.60	.043
				Consequences-Remote	.45	.132
				Consequences-Obvious	−.33	.271

[a] College attended and the five Formulating Hypotheses (FH) pretest scores used as covariates.
[b] College attended used as the only covariate.

cal variate; this information makes clear what constitutes each canonical variate and thus which dependent variables contribute most to the means shown in the preceding column. The last column shows the *p* values for univariate tests for the dependent variables shown in the preceding column. An entry was made in the last column if (*a*) $r \geq .30$ for the canonical variate or (*b*) $p \leq .05$ for the univariate significance level.

Results for treatment, sex, ideational fluency, and test anxiety. The largest *R* in Table 6 is .46 for ideational fluency, whose relationship to the first canonical variate was highly significant ($p < .001$). The correlations of dependent variables with the canonical variate show that ideational fluency was related mainly to Consequences-Obvious but also to Consequences-Remote and Formulating Hypotheses Scores 1 and (to some extent) 2. Univariate tests were significant for all these variables. The results support Guilford's (1967) placement of both the Theme test and Consequences-Obvious in the same cell of the structure-of-intellect model, and they also show that the number of hypotheses score has a large component of divergent production of semantic units.

The second-largest *R* is .37, for treatments. The five Formulating Hypotheses pretest scores as well as college attended were used as covariates for this design factor. (The effects of treatments were greater—$R = .37$ as compared with .31—and more clearly focused when these six covariates were used than when only college attended was used as a covariate, while the pattern of performance remained the same.) Quantity treatment improved performance as measured by Formulating Hypotheses Score 1 (number of hypotheses) and Formulating

TABLE 8

RESULTS OF THE MULTIVARIATE ANALYSIS OF COVARIANCE FOR TREATMENT, SEX, VOCABULARY, AND TEST ANXIETY

Effect	Multivariate significance level	M		Correlation with canonical variate		p (for univariate tests)
Treatment[a]	$F = 4.162$ $p < .001$ $R = .363$	Control group Quality treatment Quantity treatment	.01 −.46 .45	FH Score 1 FH Score 2 FH Score 5 FH Score 3	.80 .48 −.48 −.26	.001 .002 .002 .027
Sex[b]	$F = 4.742$ $p < .001$ $R = .294$	Male Female	−.30 .30	Consequences-Obvious FH Score 2 FH Score 1 FH Score 5	.67 .46 .40 .34	.001 .008 .020 .049
Vocabulary[b]	$F = 5.128$ $p < .001$ $R = .304$	Low High	−.31 .31	FH Score 3 FH Score 4 FH Score 2 FH Score 5	.78 .65 .58 .46	.001 .001 .001 .005
Test anxiety[b]	$F = 1.415$ $p < .140$ $R = .214$	Low Middle High	.25 −.02 −.23	Consequences-Remote FH Score 2 FH Score 1 FH Score 4	.59 .42 .32 −.30	.032 .219 .324 .336
Sex × Vocabulary interaction[b]	$F = 2.498$ $p < .016$ $R = .218$			Consequences-Obvious FH Score 2 FH Score 3 FH Score 4	−.53 .52 .51 .50	.026 .029 .033 .037
Treatment × Test Anxiety interaction[a]	$F = 1.261$ $p < .164$ $R = .216$			FH Score 4	—	.012

[a] College attended and the five Formulating Hypotheses (FH) pretest scores used as covariates.
[b] College attended used as the only covariate.

Hypotheses Score 2 (number of acceptable hypotheses), while quality treatment tended to produce fewer, longer, and somewhat better responses. The univariate tests for the four Formulating Hypotheses scores were all significant, including those for Formulating Hypotheses Scores 3 and 5. A one-way analysis of variance, with Formulating Hypotheses Score 3 as the dependent variable and with college attended and Formulating Hypotheses Pretest Score 3 as covariates, showed that mean performance of the quality treatment group on Formulating Hypotheses Score 3 was significantly higher than that of the control group ($F = 7.34$, $p < .007$). Thus both quality and quantity treatments produced the expected changes in performance.

The R for sex is. 26. Females were superior to males on a canonical variate that is positively correlated with Consequences-Obvious and Formulating Hypotheses Scores 2 and 5 and negatively correlated with Consequences-Remote. Thus females were found to be superior with regard to performance on tests that reflect number of hypotheses, number of words, and number of obvious consequences, but they were poorer on remote consequences.

The multivariate F for test anxiety was not significant ($p < .12$), although the univariate test for Consequences-Remote was significant ($p < .03$).

The multivariate analysis of covariance test for a Treatment × Anxiety interaction yielded a nonsignificant F. However, the univariate test for Formulating Hypotheses Score 4 was significant ($p < .014$). Exami-

nation of the nine cell means for Formulating Hypotheses Score 4 shows a tendency for low-anxious subjects in the treatment groups to outperform low-anxious subjects in the control group, while the opposite was true for high- and middle-anxious subjects. Thus either quantity or quality treatment appears to benefit low-anxious subjects more than middle- or high-anxious subjects.

Results for treatment, sex, ideational fluency, and vocabulary. Table 7 reports the multivariate analysis of covariance results for four design variables that include vocabulary. Verbal ability, as measured by the Vocabulary test score, was significantly related ($R = .30$, $p < .001$) to a canonical variate that clearly reflects quality of performance on Formulating Hypotheses and a tendency to write long responses. The univariate tests were all significant, and the relationship to verbal ability was positive, as is shown by the means.

The results for treatments, sex, and ideational fluency were all quite similar to those shown in Table 6, except that in the case of sex the univariate tests were not significant for Formulating Hypotheses Scores 2 and 5. Inclusion of vocabulary as a design variable has apparently removed some of the variance that was attributed to sex in the analyses reported in Table 6.

Univariate tests show evidence of an interaction of sex and vocabulary involving Formulating Hypotheses Scores 2, 3, and 4, the measures reflecting quality of performance on Formulating Hypotheses. The means on Formulating Hypotheses Score 2 in the 2 × 2 interaction table show that females generally earned higher scores, but the difference between males of high and low verbal ability was much greater than that between females of high and low verbal ability.

Results for treatment, sex, vocabulary, and test anxiety. Table 8 presents results for the third combination of design variables, which provides an opportunity to see if there is a Vocabulary × Anxiety interaction. None was found. The Sex × Vocabulary interaction did appear again, and in this analysis the multivariate test was significant ($p < .016$) as were the univariate tests. The univariate test for the Treatment × Anxiety interaction was again significant for Formulating Hypotheses Score 4. Generally speaking, results for main effects were very similar to those found in the other two analyses except for some differences in details of the canonical variate for sex, which are attributable to the variations in design factors.

Discussion

The effects of treatment on change in performance were shown to be highly significant; the proportion of variance in the canonical variate accounted for by treatment is about .13 when pretest scores are used as covariates. The effect of quantity treatment was basically to increase the number of Formulating Hypotheses responses and decrease the average number of words per response, while the effect of the quality treatment was to increase the average number of words per response, increase the rated quality of responses, and decrease the number of responses. The possibility of improving quality, as well as quantity, of creative performance by using models is thus demonstrated. The effect of the quality treatment is smaller, as judged by correlations of Formulating Hypotheses scores with the canonical variate; but a separate one-way analysis of variance confirms the finding that the quality models result in responses of higher quality. The change in performance tends toward a literal copy of the models in terms of number, length, and quality of responses. No evidence of transfer of training to the Consequences test was found.

Ideational fluency was found to account for a relatively large proportion of the variance in the domain of dependent variables ($R^2 \cong .23$), and the other ability measure employed, the Vocabulary test, was also significantly related to performance ($R^2 \cong .09$). These two ability measures predict quite different aspects of performance: Ideational fluency is related to quantity of production (especially the Consequences test scores and Formulating Hypotheses Score 1), while vocabulary is related to quality of performance. A definition of creativity in

terms of fluency would appear to be correct only if the quality of creative performance were ignored.

Sex was also found to be a significant factor ($R^2 \leqq .07$), although the proportion of variance contributed depends somewhat on which other design factors are included in the analysis (since sex is significantly correlated with anxiety, vocabulary, and especially, ideational fluency). Females were generally superior on a canonical variate that correlates positively with Consequences-Obvious and Formulating Hypotheses Scores 2 (number of acceptable hypotheses) and 5 (average number of words).

Test anxiety accounted for an insignificant proportion of the variance ($R^2 \leqq .04$); high anxiety was associated with poorer performance on Consequences-Remote. None of the univariate tests were significant for any Formulating Hypotheses score, and there was no evidence of a nonlinear relationship to the canonical variate.

A salient finding of the earlier study was a U-shaped relationship between test anxiety and number of hypotheses. A possible reason for the failure to replicate the curvilinear relationship is that the relationship of performance to anxiety is different for males and females. (The previous study used only male subjects.) Although a significant Anxiety × Sex interaction was not found, a plot (not shown) of Formulating Hypotheses Score 1 (the variable involved in the earlier study) against the three levels of test anxiety showed that the same U-shaped curve existed for male subjects, while for females the curve was more or less linear and descending (high anxiety associated with fewer hypotheses). Thus the data at least suggest that the relationship for males is similar to that found previously and that sex differences exist.

Weak evidence of a Treatment × Anxiety interaction was found involving Formulating Hypotheses Score 4 (average scale value). Since the hypothesis that motivated the original study (Klein et al., 1969) was that anxious individuals would profit more from the treatments than less anxious people, and that hypothesis was not then confirmed, the finding of even a weak Treatment × Anxiety interaction is of interest—even though Formulating Hypotheses Score 4 was involved rather than Formulating Hypotheses Score 1. A plot (not shown) of treatments against Formulating Hypotheses Score 4 means for the three levels of test anxiety showed relationships completely unlike those predicted for Formulating Hypotheses Score 1. The plot shows that for the control group, quality of performance was poorest for low-anxiety individuals, while for both treatment groups performance was higher for low-anxiety subjects. At higher levels of anxiety, the performance of control group members was superior to that of the treatment groups. Thus the relationship was the opposite of what had been predicted for Formulating Hypotheses Score 1. However, since Formulating Hypotheses Score 1 is a quantity score and Formulating Hypotheses Score 4 a quality score, the results may not be entirely inconsistent with the original hypothesis.

A weak Sex × Vocabulary interaction was also found, the correlates of the canonical variate including quality measures (Formulating Hypotheses Scores 2, 3, and 4) and Consequences-Remote. Examination of appropriate plots showed that performance of females was generally superior, but there was relatively little difference between males and females of high ability while low-ability females were much superior to low-ability males.

Formulating Hypotheses appears to possess appropriate psychometric properties for further explorations in the realm of creative performance. It possesses a certain amount of face validity, the items being concerned with interpretation of real data obtained in various kinds of scientific undertakings, and therefore may possess certain advantages over such tests as "brick uses" and "consequences." The scores so far developed are reasonably adequate from the standpoint of reliability, and the interitem correlations are sufficiently high that one could build a test of almost any reliability he desires by increasing the number of items. The span of abilities covered by the present five scores appears to include quantity of performance, quality of performance, and length of re-

sponses. It would be desirable to add scores measuring rarity or originality of responses. The study provides some evidence of the construct validity of the test, since the scores generally relate to other measures and to treatments in ways that are logical or in accordance with theoretical expectations. The use of tests like Formulating Hypotheses may be useful as provisional criterion measures in investigations of scientific creativity—its trainability, influences of situational factors on it, and the cognitive, attitudinal, and temperamental characteristics associated with it.

The results were, in general, consistent with our armchair analysis of the processes involved in formulating hypotheses. The Formulating Hypotheses quantity scores were related to fluency and the quality scores to verbal ability (as measured by a vocabulary test). The relationships of Formulating Hypotheses scores to the experimental treatments were significant and in the expected directions. These relationships may be interpreted as showing that variations in creative performance may be influenced by altering subjects' standards as to what constitutes satisfactory performance. (Another possible interpretation is that the treatments changed ability, but the more parsimonious and intuitively appealing hypothesis is that presenting models of satisfactory performance altered standards rather than ability. It is unlikely that cognitive abilities can be substantially changed by a small amount of training that merely involves presentation of models.)

The lack of significant relationships of Formulating Hypotheses scores to test anxiety scores makes less tenable the notion that amount of self-censorship of ideas is a function of anxiety.

REFERENCES

ALPERT, R., & HABER, R. M. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, 1960, **61**, 207–215.

CLYDE, D. J., CRAMER, E. M., & SHERRIN, R. J. *Multivariate statistical programs*. Coral Gables, Fla.: Biometric Laboratory of the University of Miami, 1966.

FREDERIKSEN, N. *Development of the test "Formulating Hypotheses": A progress report*. (Office of Naval Research Technical Report, Contract Nonr-2338(00)) Princeton, N. J.: Educational Testing Service, June 1959.

FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. *Kit of Reference Tests for Cognitive Factors*. Princeton, N. J.: Educational Testing Service, 1963.

GUILFORD, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1967.

HARLESTON, B. W. Test anxiety and performance in problem-solving situations. *Journal of Personality*, 1962, **30**, 557–573.

KLEIN, S. P., FREDERIKSEN, N., & EVANS, F. P. Anxiety and learning to formulate hypotheses. *Journal of Educational Psychology*, 1969, **60**, 465–475.

LEVY, L. H. Originality as role-defined behavior. *Journal of Personality and Social Psychology*, 1968, **9**, 72–78.

STENNETT, R. G. The relationship of performance level to level of arousal. *Journal of Experimental Psychology*, 1957, **54**, 54–61.

STRATTON, R. P., & BROWN, R. Improving creative thinking by training in the production and/or judgment of solutions. *Journal of Educational Psychology*, 1972, **63**, 390–397.

(Received January 15, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 83-89

ELABORATIVE PROMPT EFFECTS IN CHILDREN'S PAIRED-ASSOCIATE LEARNING: DESIGN AND POPULATION COMPARISONS[1]

JANET D. FREDERIKSEN AND WILLIAM D. ROHWER, JR.[2]

University of California, Berkeley

The effects of elaborative prompts on noun-pair learning in children were examined. The effects of these prompts were compared under two experimental designs—between subjects or within subjects. Of the total sample of 194 third-grade children, half were drawn from a middle-class and half from a lower-class population. Every child was administered perceptual and verbal ability tasks as well as a paired-associate task consisting of a list of 25 noun pairs presented in one of seven different ways. The results revealed that estimates of prompt effects vary with both design type and with population. Correlations among performances measured on the various tasks were generally higher for the lower-class than for the middle-class samples.

A general finding has been that middle-class children, on the average, obtain higher scores on standard IQ tests than do lower-class children (e.g., Janke & Havighurst, 1945; Schaie, 1958). In contrast, there is evidence that such social class differences do not result when other tasks are used to index learning ability. For example, Semler and Iscoe (1963) compared the performance of black and white subjects across age levels of five through nine years. While they found significantly lower full scale Wechsler Intelligence Scale for Children (WISC) IQs for the black sample, they found that on paired-associate tasks, the performance of the white children was superior to that of the black children only at the five- and six-year-old age levels.

In an attempt to explicate this finding, Rohwer, Lynch, Levin, and Suzuki (1968) tested first-, third-, and sixth-grade children from high- and low-stratum schools on one or another of four versions of a paired-associate task. The study was specifically concerned with conditions that had been found to facilitate paired-associate learning in young children and with the effects of these conditions on social class differences. It was found that the use of connecting sentences between noun pairs, as compared with the nouns alone or the use of connecting phrases between the noun pairs, facilitated learning for all subjects. Similarly, the use of action pictures, as compared to still pictures, led to better performance. The amount learned by older subjects was greater than that learned by younger subjects regardless of school stratum. In all conditions, low-stratum children learned as efficiently as high-stratum children.

A second approach, by means of a within-subjects design, has also been used to examine the effects of learning conditions in connection with social class differences. Rohwer, Ammon, Suzuki, and Levin (1971)

[1] This research was supported in part by National Institutes of Health Grant HD03869. The report was prepared at the Institute of Human Learning, which is supported by grants from the National Institutes of Health.

The authors wish to acknowledge the assistance of Ron Karlsberg in collecting the data for the study and the cooperation of the principals, teachers, and children of the Peres, Lake, El Monte, Del Mar, and Madera schools in making the study possible.

[2] Requests for reprints should be sent to William D. Rohwer, Jr., Institute of Human Learning, University of California, Berkeley, California 94720.

selected samples of low-socioeconomic-status black and high-socioeconomic-status white children at the kindergarten, first- and third-grade levels. All subjects were administered the Peabody Picture Vocabulary Test, Children's Progressive Matrices, and a series of paired-associate-learning tasks. Each paired-associate list was composed of five item types representing different amounts of prompting for elaboration. In general, the results revealed large population differences on both IQ tests at all grade levels (the largest difference was between the third-grade samples), but the paired-associate measures summed over item types revealed a population difference only at the kindergarten level. Performance on the paired-associate items improved as a function of elaborative prompt conditions. Population differences in paired-associate learning varied across prompt conditions. Such differences were larger when nouns were presented alone or when still pictures were presented alone than when nouns and still pictures were presented together. In fact, the mean differences on nouns plus still picture items in both first grade and in third grade favored the low-socioeconomic-status black samples.

Despite the apparent similarity of outcome revealed by the between- and within-subjects designs, a difficulty arises when one wishes to generalize solely from the mixed-list results. Mallory (1972) has suggested that an estimate of the relative effect of a given prompt condition based only on data from a mixed list may be misleading. He found that when a mixed list was used, second-grade children recalled more items of a particular type (when the remaining items were presented with elaborative prompts) than they did when the same item type was presented in a pure list. He further suggested that the mixed list may serve to differentiate children in terms of the strategies they employ in learning a list; for example, some children may be visualizers and tend to concentrate on the visually prompted (action) items, while others may be verbalizers and tend to concentrate on the verbally prompted (sentence) items. Accordingly, one purpose of the present study was to clarify the issues raised by the use of a within-subjects design to investigate elaborative prompt effects and population differences.

A second purpose of the present research was to gain more information about the relationship between standard ability measures and performance on a paired-associate task. There is evidence from studies of adults that the magnitude of correlation between performance on memory tasks and performance on specific ability tests is dependent upon the characteristics of the memory task (Frederiksen, 1969). In the present study it is predicted that larger correlations would be observed between congruent learning and ability measures than between incongruent ones. Specifically, larger correlations were expected between verbal tests and verbal paired-associate items and between a perceptual test and visual paired-associate items than for the remaining combinations of measures.

The ability measures were also of concern in connection with the issue of population differences. Previous evidence suggests that there is a greater degree of differentiation of abilities in middle-class school-age children than in lower-class school-age children (e.g., Dockrell, 1965; Mitchell, 1956). In the present study, it was predicted that the intercorrelations among the ability measures and that the correlations between the ability measures and paired-associate learning would be greater for the lower-class children than for the middle-class children. This prediction was derived from the hypothesis that the lower degree of differentiation previously observed in lower-class children is due to the relatively limited range of cognitive strategies they employ.

This hypothesis is consistent with the position espoused by Ferguson (1954) and developed by Frederiksen (1969). It is assumed that middle-class children have more opportunities to work on tasks that are in some ways similar to the tasks in the present study. Thus, they would be more likely to have developed specific strategies for selective use in varying situations. Because of this, the intercorrelations of their performances on the various tasks should be lower than those of the lower-class children.

Method

Subjects

One hundred and ninety-four third-grade children served as subjects. The children came from five schools in a local school district, a district serving several diverse communities. On the basis of available census tract information, two schools that served lower-class areas were selected, and three schools that served middle- to upper-middle-class areas were selected. Ninety-seven children were randomly chosen from each of the two populations. Of the children in the lower-class sample, 60 were black, 32 were from white English-speaking families, and 5 were of Spanish-American descent but were fluent speakers of English. Of the 97 middle-class children, 74 were white, 13 were black, and 10 were Oriental. In the lower-class sample, 72 parental occupations could be rated on the Hollingshead scale (Hollingshead & Redlich, 1958). Of these 72, all fell within or below the fourth category (owners of small business, clerical and sales workers, and technicians) with only 11 occupations falling within that particular category. In the middle-class sample, 94 parental occupations could be rated on the Hollingshead scale. Of these, 38 fell within the first category (executives and proprietors of large concerns and major professionals) and all but 18 fell within the first three categories.

The mean age of the lower-class sample was 9.0 years with a standard deviation of 5.45 months; the mean age of the middle-class sample was 8.96 years with a standard deviation of 4.77 months.

Design

Ability tests. Children from the two social class groups were given two parts of the Primary Mental Abilities for Grades 2–4 (Thurstone, 1963), the perceptual speed test and the verbal meaning test. The verbal meaning test consists of two parts, vocabulary and sentence completion.

Paired-associate-learning task. One group of children from each social class group was tested under a mixed-list condition; that is, they were given a paired-associate list containing items representing five conditions that differed in the amount and type of elaborative prompting presented within the item. Six other groups from each social class were each tested under one of six treatment conditions consisting of a paired-associate list containing items, all of which represented one particular condition of elaborative prompting.

Materials

Each paired-associate list consisted of the same 25 pairs. The 25-item mixed list was composed of 5 item types of 5 pairs each, chosen to be consistent with Rohwer et al. (1971). The types represented the following methods of presentation: (*a*) nouns (spoken nouns, no pictures); (*b*) still pictures (no verbalization); (*c*) nouns–still (a combination of *a* and *b*); (*d*) sentence–still (the two nouns of the paired-associate sequence formed the subject and object of an action verb; for example, the *celery* climbed the *stairs*); and (*e*) nouns–action (in action sequences, the objects were shown involved in the actions described by the sentences; for example, the celery was shown moving up the stairs). The order of the pairs in the list was random with respect to item type with the restriction that all types be represented once in each sequence of five pairs.

The other six conditions (the independent group conditions) involved the presentation of one of six pure lists; that is, each list was composed exclusively of items of a single type. Five of these conditions consisted of the five methods of presentation employed in the mixed-list condition. The additional independent group was a sentence–action condition, a condition that combined the two elaborative prompting techniques of sentence descriptions and action-picture realizations of those descriptions.

The paired-associate materials were recorded on videotape. All verbalization was presented in a male voice.

Procedure

Fourteen subjects from each of the social class groups were tested under each of the six independent treatment conditions, and 13 under the mixed-list condition.

All children were initially given two parts of the Primary Mental Abilities tests. The tests were administered in individual classrooms with approximately 30 children per class. The perceptual speed test was administered prior to the administration of the verbal meaning test.

The paired-associate task was administered individually. In all conditions the child was seated in front of a television monitor. Subjects were told that they would be asked to provide the name of the missing object when shown or told the other object that had originally appeared with it. Four example pairs were presented on the screen, followed by a test trial. Instructions and examples were repeated if the subject did not comprehend the task. The appropriate 25-item list then followed. The lists were presented for two study trials, each of which was followed by a test trial. Presentation time per stimulus pair was four seconds with a one-second pause between pairs. The same rate was used for the presentation of the individual items in the test trial.

All testing was done by one of two experimenters, both of whom were white, one male and one female. The experimenters were approximately balanced with regard to schools, teachers, and experimental conditions.

Results

Social Class Effects on Ability and Learning Tasks

The mean ability test results are found in Table 1. The two parts of the verbal mean-

TABLE 1
MEAN PERFORMANCE ON ABILITY TESTS AS A FUNCTION OF POPULATIONS

Social class	Ability test		
	Vocabulary	Sentence completion	Perceptual speed
Lower	20.60	19.45	23.43
Middle	26.08	24.21	28.14

ing test are given separately as vocabulary and sentence completion. The correlation between these two test parts, for the entire sample, was .60. It was decided that a correlation of this magnitude did not justify combining the two parts into one test score.

An analysis of variance of the three ability test measures with the seven treatment conditions nested within social class was performed. Treatment conditions were included to ascertain whether there were sampling differences with respect to the ability measures for the different treatment groups. This analysis yielded an overall significant social class effect (multivariate $F = 38.79$, $df = 3/178$, $p < .0001$). The univariate tests revealed a significant social class effect for each of the three measures (vocabulary, $F = 96.86$, $df = 1/180$; sentence completion, $F = 90.00$, $df = 1/180$; perceptual speed, $F = 21.73$, $df = 1/180$, all $ps < .0001$). There were no significant treatment effects within either social class; thus, there is no evidence of selective sampling across the treatment conditions.

Mean paired-associate-learning-task results are found in Table 2. An analysis of variance of the paired-associate-learning data for the six pure-list conditions was performed with social class nested within treatment conditions. By multivariate test (in which the two trials were treated as variables) the effect of social class was significant (multivariate $F = 2.36$, $df = 12/320$, $p < .0066$). Results of the associated univariate tests, however, indicated that the difference was almost entirely located on Trial 2 (step down $F = 3.49$, $df = 6/156$, $p < .0030$) with no significant difference on Trial 1 ($F = 1.28$, $df = 6/156$). A further breakdown of this analysis revealed only one treatment condition for which there was a significant social class effect. This was the nouns condition (multivariate $F = 4.36$, $df = 2/155$, $p < .0145$). Again the locus of the effect was the second trial.

A multivariate analysis of the mixed-list-learning data (10 variables—five item types on each of two trials) again revealed a significant overall social class effect (multivariate $F = 2.72$, $df = 10/15$, $p < .0391$), thus showing that the pattern of results was different for the two social class groups. Only one univariate test was significant, however. This was for the sentence–still items on the second trial.

Condition Effects and Design Comparisons in Paired-Associate Learning

A graph of the mean paired-associate-learning performance for both the independent groups and the mixed-list condition can be found in Figure 1.

To analyze the effects of prompt conditions, the simple effects of conditions were assessed within social class. For the between-subjects design, the following contrasts between group means were performed separately for each social class: (*a*) nouns versus still; (*b*) nouns–still versus still; (*c*) nouns–action versus nouns–still; (*d*) sentence–still versus nouns–still; and (*e*) sentence–action versus nouns–action. Tests corresponding to the first four of these contrasts were also

TABLE 2
MEAN PERFORMANCE IN PAIRED-ASSOCIATE LEARNING AS A FUNCTION OF POPULATIONS AND CONDITIONS

Condition	Social class	Paired-associate task	
		Trial 1	Trial 2
Nouns	Lower	3.86	6.00
	Middle	3.79	8.64
Still	Lower	7.64	13.93
	Middle	8.86	13.07
Nouns–still	Lower	9.00	15.50
	Middle	10.50	18.14
Sentence–still	Lower	9.79	16.79
	Middle	13.07	19.71
Nouns–action	Lower	13.79	19.71
	Middle	14.00	21.71
Sentence–action	Lower	15.50	21.14
	Middle	13.57	20.64
Mixed list	Lower	11.15	16.54
	Middle	11.08	18.15

made for the mixed-list design. These tests involved the formation of four new variables representing the differences between item types and the performance of an exact test on each of these variables to determine if it differed significantly from zero (i.e., $H_0: \mu = 0$).

A summary of the results can be found in Table 3. For the six independent groups, wherever item type differences were found, they were found in both social class groups. Pictorial presentation was found to yield superior performance to verbal presentation (still > nouns), with the addition of labels to the pictures yielding even better performance (nouns–still > still). Increased prompting for visual elaboration was also found to improve performance (nouns–action > nouns–still). However, in contrast to previous findings (Rohwer et al., 1968), increased prompting for verbal elaboration did not significantly affect performance

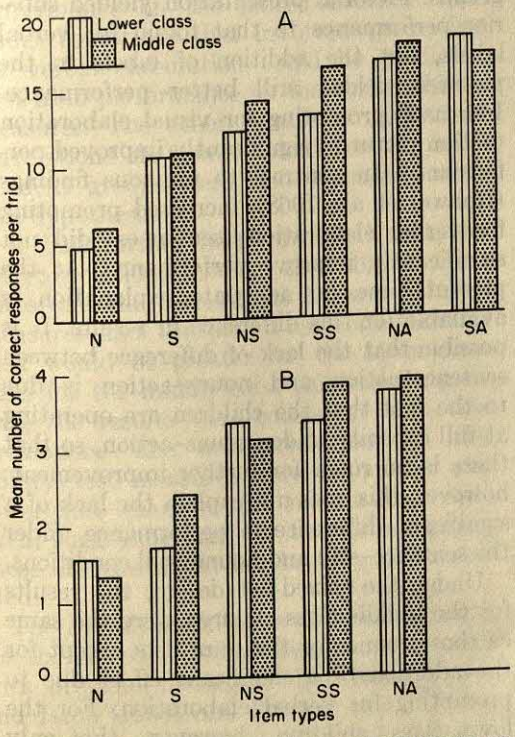


FIG. 1. Mean number of correct responses on the paired-associate task as a function of populations, conditions, and designs. (Upper panel: between-subjects design; lower panel: within-subjects design. Abbreviations: N = nouns; S = still; NS = nouns–still; SS = sentence–still; NA = nouns–action; SA = sentence–action.)

TABLE 3

CONDITION EFFECTS AS A FUNCTION OF POPULATIONS AND DESIGNS

Source	Social class	Between-subjects F (df = 1/156)	Within-subjects F (df = 1/24)
Nouns vs. still	Lower	103.77**	.14
	Middle	12.28**	7.01*
Nouns–still vs. still	Lower	32.67**	24.79**
	Middle	6.14**	4.55*
Nouns–action vs. nouns–still	Lower	19.74**	3.45
	Middle	6.81**	13.79**
Sentence–still vs. nouns–still	Lower	.58	.00
	Middle	2.34	7.47*
Sentence–action vs. nouns–action	Lower	1.34	
	Middle	.31	

* $p < .05$.
** $p < .01$.

(sentence–still vs. nouns–still, sentence–action vs. nouns–action).

For the mixed list, the item type differences are not consistent across social class. For the middle-class subjects, all comparisons made between item types were significant, including a significant effect due to prompting for verbal elaboration. For the middle-class children, still pictures were superior to nouns alone and the addition of nouns to still pictures yielded better performance than did the pictures alone. Both prompting for verbal elaboration and prompting for visual elaboration significantly improved performance. In contrast, the only significant effect found for the lower-class group under the mixed-list design was for nouns–still versus still. As can be seen in Figure 1, the lower-class children's performance is virtually identical on the nouns–still, sentence–still, and nouns–action items. Performance on these items is far superior to that for either the nouns or the still items.

The pattern of results suggests that the effects of the prompt conditions under the between- and within-subjects design are very similar for the middle-class subjects, but that this is not the case for the lower-class subjects.

Interrelation of Paired-Associate and Ability Test Measures

As was predicted, higher correlations of the ability test measures emerged for the

lower-class group than for the middle-class group. Perceptual speed is significantly correlated with the two verbal measures for the lower-class sample (vocabulary, $r = .39$; sentence completion, $r = .44$, $n = 97$, $p < .01$); neither of these correlations was significant for the middle-class sample. The correlation of the two verbal measures is .56 for the lower-class sample and .48 for the middle-class sample ($n = 97$, $p < .01$).

The correlations between the individual ability test and paired-associate learning for each prompting condition are interesting, although they must be interpreted with caution because of the small sample sizes on which they are based. For the pure lists for the lower-class sample, perceptual speed correlated significantly with Trial 1 performance under the still condition ($r = .55$, $n = 14$, $p < .05$). Significant correlations were also found for this sample between both verbal ability measures and Trial 1 sentence–still performance (vocabulary, $r = .59$; sentence completion, $r = .60$, $n = 14$, $p < .05$) and between perceptual speed and Trial 1 of the sentence–action condition ($r = .74$, $n = 14$, $p < .05$). Of the significant correlations found for the middle-class sample, none of these involved perceptual speed. The significant correlations that were found were those of vocabulary with sentence–still performance on Trial 1 ($r = .53$, $n = 14$, $p < .05$) and of both verbal ability measures with sentence–action performance (Trial 1, sentence completion, $r = .60$; Trial 2, vocabulary, $r = .55$; sentence completion, $r = .62$, $n = 14$, $p < .05$). It should be noted here that for the middle-class group, performance on the paired-associate task under the conditions involving prompting for verbal elaboration can be predicted from performance on the verbal ability measures. This relationship holds for the sentence–still but not for the sentence–action condition for the lower-class group.

Very few significant correlations were found under the mixed-list condition. In the lower class, the significant correlations were perceptual speed with the still condition, Trial 1 ($r = .57$, $n = 13$, $p < .05$); vocabulary with nouns, Trial 2 ($r = .57$, $n = 13$, $p < .05$); and sentence completion with sentence–still, Trial 2 ($r = -.68$, $n = 13$, $p < .05$). In the middle class, the significant correlations were perceptual speed with nouns, Trial 1 ($r = -.62$, $n = 13$, $p < .05$) and vocabulary with nouns–action, Trial 1 ($r = .59$, $n = 13$, $p < .05$).

Discussion

The principal aim of this study was to compare a within-subjects (mixed-list) with a between-subjects (pure-list) design, using the same items and conditions in both. The question of concern was that of the generalizability of results found using the mixed-list design. The general conclusion that can be reached is that these two designs yield similar results for the middle-class subjects but produce different outcomes for the lower-class subjects.

The effects of elaborative prompting in the between-subjects design showed the same pattern of results for both social class groups. Pictorial presentation yielded superior performance to that found for verbal labels, but the addition of labels to the pictures yielded still better performance. Increased prompting for visual elaboration (action pictures) significantly improved performance. In contrast to previous findings (Rohwer et al., 1968), increased prompting for verbal elaboration (sentences) did not significantly improve performance. At the present time no adequate explanation is available for this difference in results. It is possible that the lack of difference between sentence–action and nouns–action is due to the fact that the children are operating at full capacity under nouns–action, so that there is no room for further improvement; however, this does not explain the lack of a significant difference in performance under the sentence–still and nouns–still conditions.

Under the mixed-list design, the results for the middle-class children were the same as those found for the pure lists except for the addition of a significant effect due to prompting for verbal elaboration. For the lower-class children, however, the only comparison that yielded a significant effect was that of nouns–still with still. It appears that for these children, with the mixed-list design, performance is approximately equal under all conditions involving

elaborative prompting and that this performance far exceeds that found for either the simple use of nouns or still pictures. It appears as if these children are selecting for attention the items containing elaborative prompting. These items would prove much easier for the lower-class children if, as Rohwer (1968) suggests, they do not spontaneously employ elaboration but do utilize elaboration prompts.

A second focus of the study was on the relationship between standard ability measures and the paired-associate-learning task. It had been predicted that the pattern of correlations between the two sets of measures would be dependent on task characteristics, specifically, that the largest correlations would be found between the verbal tests and verbal items and between the perceptual test and pictorial items. Some support for this hypothesis can be seen in the finding that for the independent groups, performance on the paired-associate task under conditions that involved prompting for verbal elaboration could be predicted best from performance on the verbal ability measures. The other significant individual correlations that were found for the independent groups were also consistent with this hypothesis.

It had further been predicted that the intercorrelations among the ability measures and the correlations between ability measures and the paired-associate-learning task would be greater for the lower-class than for the middle-class children. In general, the results were consistent with this prediction. The intercorrelations of the ability test measures were higher for the lower-class than for the middle-class children, suggesting that for the former group these abilities might not be very differentiated.

The overall social class differences found for the paired-associate tasks are congruent with previous findings. While both Semler and Iscoe (1963) and Rohwer et al. (1968) reported no social class differences, neither of these studies employed the particular condition that in the present study showed a significant social class effect, that is, the nouns condition. This result is consistent with that previously reported for third-grade samples by Rohwer et al. (1971), using a mixed-list design.

REFERENCES

Dockrell, W. B. Cultural and educational influences on the differentiation of ability. In, *Proceedings of the 73rd Annual Convention of the American Psychological Association 1965.* Washington, D. C.: APA, 1965. (Summary)

Ferguson, G. A. On learning and human ability. *Canadian Journal of Psychology,* 1954, **8,** 95–112.

Frederiksen, C. H. Abilities, transfer, and information retrieval in verbal learning. *Multivariate Behavioral Research Monographs,* 1969, No. 69-2.

Hollingshead, A. B., & Redlich, F. C. *Social class and mental illness: A community study.* New York: Wiley, 1958.

Janke, L. L., & Havighurst, R. J. Relations between ability and social status in a midwestern community: II. Sixteen-year-old boys and girls. *Journal of Educational Psychology,* 1945, **36,** 499–509.

Mallory, W. A. Abilities and developmental changes in elaborative strategies in paired-associate learning of young children. *Journal of Educational Psychology,* 1972, **63,** 202–217.

Mitchell, J. V., Jr. A comparison of the factorial structure of cognitive functions for a high- and low-status group. *Journal of Educational Psychology,* 1956, **47,** 397–414.

Rohwer, W. D., Jr. Mental mnemonics in early learning. *Teachers College Record,* 1968, **70,** 213–226.

Rohwer, W. D., Jr., Ammon, M. S., Suzuki, N., & Levin, J. R. Population differences and learning proficiency. *Journal of Educational Psychology,* 1971, **62,** 1–14.

Rohwer, W. D., Jr., Lynch, S., Levin, J. R., & Suzuki, N. Grade level, school strata, and learning efficiency. *Journal of Educational Psychology,* 1968, **59,** 26–31.

Schaie, K. W. Occupational level and the primary mental abilities. *Journal of Educational Psychology,* 1958, **49,** 299–303.

Semler, I. J., & Iscoe, I. Comparative and developmental study of the learning abilities of Negro and white children under four conditions. *Journal of Educational Psychology,* 1963, **54,** 38–44.

Thurstone, T. G. *Primary Mental Abilities for Grades 2–4.* Chicago: Science Research Associates, 1963.

(Received January 22, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 90-95

MONETARY INCENTIVE EFFECTS UPON INCIDENTAL LEARNING DURING AN INSTRUCTIONAL TASK

STEPHEN WOLK[1]
University of Maryland

JOSEPH DuCETTE
Temple University

Incidental learning was analyzed in regard to attentional arousal and monetary incentive. Subjects scanned passages of a narrative for typographical and conceptual errors (intentional performance) but were tested additionally for the retention of various aspects of the story (incidental learning). Results indicated (*a*) specific instructional directions increased incidental learning without necessarily decreasing intentional performance; (*b*) less specific instructional arousal interacted with the presence or absence of a monetary incentive, suggesting the deleterious effects of too much motivation and/or cognitive support upon the learning of irrelevant task material; (*c*) this interactive effect of incentive and arousal extended into the more traditional measures of recall and recognition of facts as well as the retention of temporal sequences within the material.

The area of incidental learning remains a highly differentiated collection of research. While this work has considered a varied list of individual and situational factors and their effects upon levels of incidental learning, the proliferation of experimental designs and procedures (types of tasks, instructions to subjects and operational definitions of variables) has made a systematic evaluation of any subset of this research extremely difficult. Additionally, research along these lines considering theoretical questions and materials of interest to the educator remains somewhat scant. The purpose of the present set of experiments was to examine the effects of monetary incentives and type of instructions upon the level of incidental-type learning related to educational settings.

While the primary focus of this study concerns the effects of incentives in incidental learning, the other manipulation in this experiment (type of instruction) was chosen because of its obvious significance and because of the variance in the reported research. In other research similar to the methodology of the present study (DuCette & Wolk, 1973), it has been argued that instructions increase incidental learning by heightening the cognitive prominence of stimuli without necessarily negating the incidental nature of these stimuli, a finding congruent with previous research. The reason for this effect remains to be determined.

Several earlier studies have suggested somewhat straightforward relationships between incentives and intentional and incidental performance, such as that of Bahrick (1954) who found that monetary reward for intentional performance significantly increased this performance. However, incidental learning decreased when such an incentive was present. Johnson and Thompson (1962) employing a similar design confirmed these findings. Zeffy and Bruning (1966) have also supported Bahrick, using measures of anxiety as the motivationally produced drive.

Later research and writing in this area, thus, have come to distinguish two types of motivational effects possible in the typical incidental-learning paradigm. Congruent with the work of Bahrick and of Johnson and Thompson, there is an incentive-oriented type of drive, acting to "funnel" attention in a directive manner, resulting in a reduction of attention to the irrelevant

[1] Requests for reprints should be sent to Stephen Wolk, Institute for Child Study, College of Education, University of Maryland, College Park, Maryland 20742.

or incidental material. Additionally, however, a long list of investigations has considered the effects of an emotionally-based drive, usually represented by an individual assessment or a situational inducement of anxiety. This type of drive is interpreted as producing a generalization of attention from relevant to irrelevant aspects of a learning situation. Most notably the earlier work of Kausler, Trapp, and Brewer (1959), using the tasks of Bahrick (1954), has led to such a distinction. In this study two experiments were reported in which anxiety was either assessed individually (Manifest Anxiety Scale) or induced by differential instructions. In both experiments the high-drive group was superior to the low-drive group in intentional learning but did not differ in incidental learning, a finding contrary to that of Bahrick and his followers.

A more extensive reading of the literature, however, suggests that the plausibility of the distinction between incentive and emotionally produced motivational effects upon incidental learning may be in doubt. Duffy (1961), in an unpublished dissertation, found high-anxious subjects to remember significantly less of the incidental material, contrary to the purported generalized attention effect of such a drive, but similar to the funneling effect of an incentive-produced drive. Miller and Dost (1964) and Stanton (1971), who employed the Bahrick intentional and incidental tasks, also found detrimental effects of such a drive upon incidental learning. Indeed, Dornbush (1965) cited evidence that the presence of incentives for intentional performance may not result in a lower level of incidental learning, as Bahrick (1954) argued. These studies, contrasted with earlier ones, suggest several points:

1. The effects of an emotional drive upon incidental learning may be very similar to those of an incentive drive (i.e., funneling of attention to relevant task parameters), suggesting a single motivational effect upon such learning; additionally, heightening of the motivational state (as in the Stanton, 1971, study) may actually suppress both the levels of intentional and incidental learning.

2. The reliance on purely motivational interpretations of reduced performance under increased drive may be too simplistic. It is possible that increased activity under high-drive states may be channeled into more attention to the incidental material. This implies that heightened motivation not only changes the intensity of the subject's behavior but also his strategy and direction.

It would appear, congruent with the above points, that the literature has yet to delineate adequately the relationship between motivational states, specifically material incentives and incidental learning. Moreover, almost total disregard has been paid to an experimental task in which stimuli are interrelated as opposed to the discrete tasks discussed previously. To what degree the incidental material is intrinsic to the intentional task would appear to play an important role in determining levels of incidental learning as well as the effect of any motivational state. Being extrinsic to the main task, additionally, the incidental material may lose its "incidental nature" from a type of isolation effect. Incidental-learning research and those studies dealing with incentives have also offered little in the way of generalizability to instructional settings. With these points for consideration, two experiments were designed to determine the effects of monetary incentive upon incidental learning in conjunction with differential instructions to orient attention.

Experiment 1

Method

Subjects. The subjects were 60 students enrolled in several sections of an intermediate level educational psychology course during a summer instructional program (males = 27; females = 33). Neither course requirements nor grades were contingent upon participation in the experiment.

Materials. An intentional-incidental learning task, consisting of connected stimuli (textual material), the basis for several previous investigations (DuCette & Wolk, 1973; Wolk & DuCette, 1974), was employed. The intentional task had the subjects read a story about the discovery of a fictitious drug (written by the experimenters). The story was composed of four paragraphs of seven sentences each with approximately 200 words per paragraph. Subjects were given two minutes to read each paragraph. The general intentional instructions were that the subjects were to look

for and circle errors in the paragraphs "as if they were proofreaders." These errors were of two kinds: obvious misspellings, run-on words, duplicated letters, etc. and more conceptual errors (inappropriate words, subjects not agreeing with verbs, non sequiturs, etc.). There were 20 typographical and conceptual errors in each paragraph.

The measure of incidental learning was the retention of specific examples of the following categories of items presented in the story: numbers, colors, countries and cities, individual's names, animals, occupations, and parts of the body. A total of 36 examples of these categories were dispersed by having 9 present in each paragraph. Following the intentional reading of the story, subjects were presented first with a grid having the seven categories of elements listed. Instructions said to list as many of the examples of these categories as appeared in the story (recall measure). Subjects were allotted six minutes. Following this, subjects were presented with a list of the seven categories of elements under each of which were 15 examples. Among these were the actual words that had appeared in the story. For each category, the subject was instructed as to how many of the examples had actually appeared (this was either five or six) and was told to circle only that number for his choices but no more (recognition measure). The subject was to guess if necessary but to make sure he circled the proper number. Six minutes were also given.

Procedure and design. Subjects were randomly assigned to one of two groups: an incidental group, to which only the intentional task directions were given but for which retention of the categories was totally incidental, and an intentional group, which was given, in addition to the intentional task directions, an explicit statement concerning the nature of the incidental-learning test they would be given following the task. (This included the actual list of the seven categories of elements to be tested.) Each subject was tested in a session that lasted approximately 25 minutes.

Results

A one-way analysis was computed on three dependent variables: intentional task performance, that is, number of errors correctly identified (total possible score = 80); incidental recall of examples (total possible score = 36); and incidental recognition of examples (total possible score = 36).

Presented in Table 1 are the means and standard deviations of the three dependent variables. For intentional performance (searching for errors) there was no significant difference between groups ($F = 1.38$, $df = 1/58$). The groups did differ, however, on both measures of incidental learning: recall ($F = 16.22$, $df = 1/58$, $p < .05$) and recognition ($F = 4.06$, $df = 1/58$, $p < .05$). Implications of this will be considered in conjunction with the results of Experiment 2. Equally important, this experiment served to establish baseline rates of both intentional and incidental learning under two instructional conditions, a consideration seldom made in most incidental-learning studies.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF DEPENDENT VARIABLES, EXPERIMENT 1

Group	Intentional performance	Incidental learning	
		Recall	Recognition
Intentional			
$\bar{X}$	56.80	10.60	24.37
SD	8.93	3.50	2.69
Incidental			
$\bar{X}$	59.71	6.85	21.96
SD	9.90	3.56	3.10

EXPERIMENT 2

Method

Subjects. Subjects in this experiment were 88 students drawn from several sections of the same course as that in Experiment 1 (males = 40, females = 48). However, this occurred in a semester of the summer session subsequent to that of Experiment 1.

Materials. Both the intentional and incidental tasks were the same as those in Experiment 1, as were the measures of intentional performance, incidental recall, and recognition. However, in order to acquire a broader base of evaluation for the manipulations in this experiment, the following incidental-learning measures were added: a multiple-choice test composed of 14 items assessing retention of factual information about the story. Each stem of each question had five alternative choices from which the subject was to circle the correct one (total possible score = 14). This followed the recognition test. Lastly, subjects were given a temporal sequencing test composed of three sets of six events having occurred in the story. The subjects' task was to place these in the temporal order in which they had occurred in the story by numbering them from 1 to 6 (1 = first to have occurred). This test was scored by awarding, for each of the three sets, 2 points for each fact that had been correctly ranked and 1 point if the fact had been ranked only one step removed from its actual rank (total possible score = 36).

Procedure and design. The main experimental manipulation consisted of the use of monetary incentives for intentional task performance. In a

TABLE 2

MEANS AND STANDARD DEVIATIONS OF DEPENDENT VARIABLES BY EXPERIMENTAL GROUPS, EXPERIMENT 2

Measure	High attentional arousal		Low attentional arousal	
	No incentive	Incentive	No incentive	Incentive
Intentional performance				
$\bar{X}$	57.00	58.27	55.40	61.42
SD	10.71	10.54	10.10	8.30
Incidental recall				
$\bar{X}$	9.43	6.68	6.05	10.00
SD	3.94	2.41	3.39	3.06
Incidental recognition				
$\bar{X}$	24.23	23.04	22.30	25.18
SD	2.91	2.41	2.99	2.75
Multiple choice				
$\bar{X}$	6.63	5.13	5.00	6.64
SD	2.01	1.73	2.12	1.83
Temporal sequencing				
$\bar{X}$	18.47	16.63	15.10	18.81
SD	4.14	4.97	4.93	3.81

previous semester, students similar to the subjects in the current experiment were questioned about various types of payoffs in experiments. They were asked to judge whether individual incentives for number of errors found ($.03 per item) were less motivating than several large payoffs for the highest performers on the task. The latter alternative was judged more effective. Therefore, in the present study half of the subjects were randomly assigned to an incentive condition. When they reported for testing they were told that the three highest performers on the intentional (proofreading) task would receive prizes of $50.00, $30.00, and $10.00. Prizes were awarded at the termination of the study.

An additional manipulation consisted of attentional arousal for the incidental material through instructions. Half the subjects were randomly assigned to a condition in which no information concerning the incidental material was given, corresponding to the incidental group of Experiment 1. The remaining subjects were informed that their knowledge of "some aspects of the story" would be questioned, instructions more intermediate than those of the intentional group of Experiment 1. Thus, the influences of both material incentives and instructions, as well as the interactive influences of both, were examined for both the intentional task and the various measures of incidental learning. Subjects were tested in a 40-minute session.

Results

The statistical design was a 2 × 2 analysis of variance with independent groups. Table 2 presents the means and standard deviations of all dependent variables by experimental groups. Results of the analyses of variance revealed that there were no significant main effects or interactions for intentional performance. For all measures of incidental learning, however, the interaction term was always significant, although again there were no main effects. These interactions were obtained for the variables of recall ($F = 22.69$, $df = 1/84$, $p < .001$), recognition ($F = 11.46$, $df = 1/84$, $p < .001$), multiple choice ($F = 10.44$, $df = 1/84$, $p < .01$), and temporal sequencing ($F = 12.39$, $df = 1/84$, $p < .001$).

As a further explication of the significant interaction effects for all measures of incidental learning, comparisons among means were made using the Tukey test for differences among means (Kirk, 1968). These are summarized as follows:

1. There were no significant differences for all variables between the high-attentional–arousal–incentive and low-attentional-arousal–no-incentive conditions, nor between the high-attentional-arousal–no-incentive and low-attentional-arousal–incentive conditions.

2. Under the incentive condition, high attentional arousal produced significantly poorer incidental learning than did low attentional arousal for recall ($p < .01$), recognition ($p < .01$), multiple choice ($p < .05$), and temporal sequencing ($p < .05$).

3. Under the no-incentive condition, high attentional arousal produced significantly higher incidental learning than did low attentional arousal for recall ($p < .01$), recognition ($p < .01$), multiple choice ($p < .05$), and temporal sequencing ($p < .01$).

DISCUSSION

While it is somewhat surprising that neither attentional arousal nor incentives

had a main effect on cover task performance, it is even more interesting that there was no interaction between the two. It will be argued that the lack of these expected effects is not due to either manipulation's essential irrelevance but to the fact that cover task performance was, for all practical purposes, at asymptote for all subjects regardless of condition. This does not mean that this specific nonlearning, orienting task under the conditions used in this experiment is cognitively taxing, for we will argue later that it is not, but only that the particular conditions under which the task was given did not allow variance in performance for these particular subjects. Both experiments support this (over-all mean intentional performance for Experiment 1 = 58.25; for Experiment 2 = 58.02).

Since the pattern for all of the measures of incidental learning was the same, they are discussed together. Based on previous research it was possible to hypothesize the effects of both types of manipulations. In reference to instructions, it was logical to assume that giving subjects orienting instructions for the incidental material would enhance performance. This was strongly demonstrated in Experiment 1 in which the incidental learning was essentially intentional, but not in Experiment 2. One possible theoretical position concerning incentives, based on the work of Kausler et al. (1959) and Stanton (1971), predicts a facilitating effect of motivation due to a generalized attentional effect from the intentional to the incidental task. The other position (Bahrick, 1954; Johnson & Thompson, 1962) predicts a decrease in incidental learning due to the funneling effect on attention. It is evident that neither position is fully supported by the data. While it is true that either cue explication or incentives significantly improve performance over the baseline (the low-attentional-arousal–no-incentive condition), the two in combination seem to cancel each other out. There are two explanations of this interaction that seem plausible.

The first possibility that is explored assumes that the two manipulations used in the experiment involved two different processes—one cognitive (instructions) and one motivational (incentives). Either of these processes improves incidental learning, but for different reasons: Instructions organize and structure the input; incentives increase the attention paid to all aspects of the tasks. This could explain the equality of the high-attentional-arousal–no-incentive and low-attentional-arousal–incentive conditions. This position, however, seems incapable of explaining the low level of performance in the high-attentional-arousal–incentive condition (not different from the low-attentional-arousal–no-incentive condition). It would, in fact, predict that the condition most facilitating for incidental learning would be the condition in which both instructions and incentives were present. The data show that this prediction is not upheld, bringing such a two-process model into question, at least for this experiment.

The second possibility assumes that both cue explication and incentives are motivational manipulations. From this it follows that the reason for the poor performance in the high-attentional-arousal–incentive condition is due to the subjects being over motivated, an example of the Yerkes-Dodson law (Broadhurst, 1957). This explanation is congruent with previous theoretical interpretations of incidental learning. For example, Ryan (1970), in his review of incidental learning, points out that intentional instructions to learn (which are analogous in this experiment to the high-attentional-arousal condition) are often interpreted as being a motivational manipulation. If this is correct, then subjects in the high-attentional-arousal–incentive condition had two sources of motivation acting upon them. Since the incidental task was extremely difficult, the debilitating effect of this heightened motivation is exactly as described in previous research (Spence & Taylor, 1951). In addition, this explanation is as capable as the first in explaining the heightened performance in the two alternate cells. In both cases, the subjects were more motivated than the subjects in the low-attentional-arousal–no-incentive condition and thereby performing at greater levels. Of the two explanations, therefore, the interpretation of the results from a purely

motivational perspective seems most powerful.

The data from this experiment would seem to indicate that both the generalizing and funneling effects of motivational arousal on attention in incidental learning may be applicable, but in different situations. In the absence of other sources of motivation and/or cognitive supports, motivation may serve to generalize attention. With too much motivation or the presence of cognitive supports that place too much stress on the organism, however, attention becomes selective and narrowly focused. Further research in this area must be more precise than previous research in specifying the nature of the task used and the level of motivation employed. Since incidental learning is of such significance in all aspects of education, this point cannot be overlooked. A teacher may, in fact, be decreasing the amount of nondirected material a student learns by jointly manipulating a set of conditions under which the material is presented. When these factors and the interactions between them are delineated, the dynamics of incidental learning can be better specified.

REFERENCES

BAHRICK, H. P. Incidental learning under two incentive conditions. *Journal of Experimental Psychology*, 1954, **47**, 170–172.

BROADHURST, P. L. Emotionality and the Yerkes-Dodson law. *Journal of Experimental Psychology*, 1957, **54**, 345–352.

DUCETTE, J. P., & WOLK, S. Attentional processes in incidental learnings of connected material. Paper presented at the meeting of the Eastern Psychological Association, Washington, D. C., May 1973.

DORNBUSH, R. L. Motivation and positional cues in incidental learning. *Perceptual and Motor Skills*, 1965, **20**, 709–714.

DUFFY, J. F. The effects of induced anxiety on the incidental learning of high and low anxious subjects. Unpublished doctoral dissertation, University of Pennsylvania, 1961.

JOHNSON, R., & THOMPSON, C. Incidental and intentional learning under three conditions of motivation. *American Journal of Psychology*, 1962, **75**, 284–288.

KAUSLER, D. H., TRAPP, E. P., & BREWER, C. L. Intentional and incidental learning under high and low emotional drive levels. *Journal of Experimental Psychology*, 1959, **58**, 452–455.

KIRK, R. E. *Experimental design: Procedures for the behavioral sciences*. Belmont, Calif.: Brooks/Cole, 1968.

MILLER, M. E., & DOST, J. A. Stimulus vividness and anxiety level in intentionalincidental learning. *Psychological Reports*, 1964, **14**, 819–825.

RYAN, T. A. *Intentional behavior: An approach to human motivation*. New York: Ronald Press, 1970.

SPENCE, K. W., & TAYLOR, J. Anxiety and strength of the UCS as determiners of the amount of eyelid conditioning. *Journal of Experimental Psychology*, 1951, **42**, 183–188.

STANTON, H. E. Incidental and intentional learning—One process or two? *Australian Psychologist*, 1971, **6**, 26–30.

WOLK, S., & DUCETTE, J. Intentional performance and incidental learning as a function of personality and task dimensions. *Journal of Personality and Social Psychology*, 1974, **29**, 90–101.

ZEFFY, D. J., & BRUNING, J. L. Drive and the range of cue utilization. *Journal of Experimental Psychology*, 1966, **71**, 382–384.

(Received January 23, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 96-98

IMMEDIATE AND DELAYED RETENTION EFFECTS OF INTERSPERSING QUESTIONS IN WRITTEN INSTRUCTIONAL PASSAGES[1]

JOHN R. BOKER[2]

Pennsylvania State University

A 3 × 2 × 2 factorial design with 108 college undergraduate subjects was used to study the short- and long-term retention effects of viewing testlike factual questions in conjunction with reading a written instructional passage. A simple change in the position of the interspersed questions produced different mean postreading retention performances. Subjects in questioned groups retained significantly ($\alpha = .05$) more question-relevant information than a nonquestioned control group on both an immediate and seven-day delayed retention measure. The postquestioned group retained more question-incidental content than either the prequestioned or control groups on both retention tests.

The main purpose of this study was to investigate the effects of interspersed pre- and postquestions on the delayed retention of question-relevant and question-incidental prose material. With the exception of a few studies (Natkin & Stahler, 1969; Peeck, 1970), the focus of investigation has been on immediate retention measures obtained shortly after the subjects have read and studied the stimulus material (e.g., Bruning, 1968; Frase, 1967, 1968; Rothkopf, 1965, 1970). Although this type of short-term retention may at times be a pertinent educational objective, the goals of education more commonly involve long-term retention of previously learned material. If the delayed retention effects of interspersed questions could be demonstrated to be similar to those found in studies that used only immediate retention measures, important practical and theoretical implications for certain instructional processes could result.

[1] The author expresses his appreciation to Paul A. Games and Helen I. Snyder for their critical comments concerning the study and manuscript.

[2] Requests for reprints should be sent to John R. Boker, Department of Educational Psychology, Pennsylvania State University, 201 Social Sciences Building, University Park, Pennsylvania 16802.

METHOD

Subjects

The subjects were 108 undergraduate students enrolled in the introductory educational psychology course at Pennsylvania State University. Extra credit toward the course grade was given to every subject for participation in the study. To control for prior background knowledge, students who majored in fields concerned with biology and historical geology were not permitted to participate in the study.

Materials

Passage. A prose passage consisting of 10 sections of 250 words each was prepared from college level textbooks. Each separate section was one of a series of topically related but relatively independent factual segments covering selected topics in historical geology. The topics were chosen on the basis of being generally unfamiliar to most subjects who participated in the study, and this insured that all subjects started from the same approximate baseline in learning the prose material.

Questions. Four multiple-choice items, each with four alternatives, were constructed for each of the 10 sections. The 40 items tested relatively independent factual bits of information that were evenly spaced throughout each section. Two items for each section were selected randomly and designated as interspersed questions. These 20 items constituted a criterion posttest to measure the retention of question-relevant content. The remaining 20 items appeared in the posttest as the measure of retention of question-incidental

content. Preliminary testing indicated that both sets of questions were of equal difficulty.

Retention tests. An objective, immediate retention test was constructed by arranging the 40 multiple-choice items in random sequence. Reliability of the test (Kuder-Richardson Formula 20) was .75. A delayed retention test was constructed by rearranging the 40 items in a new random sequence. The test–retest correlation between both total test scores was .80 when computed on only the control group data over the seven-day retention interval.

Design

The design of the study implied a 3×2×2 analysis of variance with repeated measures on the last two factors. The first factor was called treatments and represented the manipulation of the position of the interspersed questions within the reading passage. The prequestioned group viewed each two-item set of interspersed questions prior to reading the prose section to which the questions related. The postquestioned group viewed the interspersed questions after reading the related section. The control group simply read through all 10 sections without viewing any interspersed questions.

The two repeated measures factors were test time and retention type. The test time factor represented the immediate and delayed retention tests; whereas retention type referred to either question-relevant or question-incidental retention.

Procedure

The subjects were randomly assigned to one of the three experimental treatments for the first part of the study. Each subject was given an envelope containing the stimulus materials and was cautioned by the experimenter not to remove the materials from the envelope until instructed to do so. The materials took the form of an unbound stack of 8½ × 11 inch mimeographed pages on which the sections of the passage and the interspersed questions (for prequestioned and postquestioned conditions) were reproduced. Each mimeographed page contained either 1 of the 10 prose sections or, when required by the experimental condition, the two interspersed questions that related to a specific prose section.

The subjects were instructed to (*a*) work through the materials at their own rate, (*b*) read each page in the sequence in which it appeared, (*c*) place each page back into the envelope after they had finished reading it, and (*d*) not go back to a page once it was placed back into the envelope. Subjects in the prequestioned and postquestioned groups received additional instructions to answer directly on the page any questions that were encountered during the reading of the passage. No knowledge of results was provided for the interspersed questions, however. These two groups were also instructed to follow the same procedure as they did with the prose section pages when each page that contained questions was completed. No note taking was allowed, and the subjects were told to expect a test upon completion of their reading.

When each subject completed his reading, he answered a short questionnaire that required some personal data and then began the immediate retention test. The subjects received the delayed retention test one week later. Both tests were intended to be power tests, therefore no time limit was imposed upon the subjects.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF RETENTION TEST SCORES FOR EACH TREATMENT GROUP

Treatment	Question relevant		Question incidental		Total	
	M	*SD*	*M*	*SD*	*M*	*SD*
Immediate retention						
Prequestions	15.19	2.23	12.22	2.23	27.41	3.75
Postquestions	16.00	2.26	14.92	2.29	30.92	3.86
Control	13.17	3.27	13.56	3.04	26.73	5.66
Delayed retention						
Prequestions	13.78	2.39	10.94	1.97	24.72	3.71
Postquestions	15.17	2.57	13.92	2.21	29.09	4.02
Control	12.53	3.03	11.47	2.46	24.00	4.70

Note. $n = 36$ for each treatment group.

RESULTS AND DISCUSSION

The mean immediate and delayed retention scores for each treatment group are shown in Table 1. The overall analysis of variance performed with respect to these means indicated a significant test time effect ($F = 63.69$, $df = 1/105$, $p < .001$). The decrease in retention means of 1.21 points (about 10%) over the one-week retention interval was expected because subjects were not permitted to review the passage at any time. However, none of the interactions involving the test time factor were significant. The absence of test time interactions indicated that the same trends were found for immediate and delayed retention measures.

The overall analysis of variance also revealed a significant Treatments × Retention Type interaction ($F = 13.37$, $df = 2/105$, $p < .001$). Analysis of the question-

relevant retention means with Tukey's WSD test ($\alpha = .05$) revealed that both the postquestion and prequestion means were significantly larger than the control mean, but they were not significantly different from each other. The means were 15.58, 14.49, and 12.85 for postquestioned, prequestioned, and control groups, respectively. The results extended previous findings that viewing interspersed pre- or postquestions during the reading of instructional passages facilitated the retention of factual content related to those questions to approximately the same degree. The present data show that this effect holds for either immediate or delayed retention measures.

A separate analysis of the question-incidental retention means (collapsed over the time factor) confirmed the previously established general superiority of the postquestioned group. Using the Tukey WSD test ($\alpha = .05$), the results indicated that the prequestion and control means were not significantly different from each other, but both were significantly lower than the postquestion mean. Means of 14.42, 11.58, and 12.51 were obtained for postquestioned, prequestioned, and control groups, respectively. Again, the results substantiated previous findings that only interspersed postquestions facilitate the retention of question-incidental factual content from written instructional passages. Prequestioning actually depressed question-incidental retention scores below those of the nonquestioned control group (although this difference was not significant). The present data extends previous findings in that the facilitative effects of postquestions on retention of incidental content were relatively constant over the seven days. The present results confirmed the overall (relevant plus incidental) superiority of interspersed postquestions in that postquestions facilitated the retention of question-relevant content equally as well as did prequestions, but the postquestions were markedly superior in their question-incidental retention effects. The mean retention differences observed in the present study give support to Rothkopf's (1970) mathemagenic hypothesis. However, if the subjects could review the text, skipping back and forth, then even postquestions might depress incidental learning from written instructional passages.

REFERENCES

Bruning, R. H. Effects of review and testlike events within the learning of prose materials. *Journal of Educational Psychology*, 1968, **59**, 16–19.

Frase, L. T. Learning from prose material: Length of passage, knowledge of results, and position of questions. *Journal of Educational Psychology*, 1967, **58**, 266–272.

Frase, L. T. Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology*, 1968, **59**, 244–249.

Natkin, G., & Stahler, E. The effects on adjunct questions on short and long-term recall of prose materials. *American Educational Research Journal*, 1969, **6**, 425–432.

Peeck, J. Effects of prequestions on delayed retention of prose material. *Journal of Educational Psychology*, 1970, **61**, 241–246.

Rothkopf, E. Z. Some theoretical and experimental approaches to problems in written instruction. In J. D. Krumboltz (Ed.), *Learning and the educational process*. Chicago: Rand McNally, 1965.

Rothkopf, E. Z. The concept of mathemagenic activities. *Review of Educational Research*, 1970, **40**, 325–336.

(Received January 29, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 99–111

INTERACTION OF LEVEL I AND LEVEL II ABILITIES WITH RACE AND SOCIOECONOMIC STATUS[1]

ARTHUR R. JENSEN[2]

University of California, Berkeley

Two levels of mental abilities were hypothesized to interact with socioeconomic status and/or race such that (*a*) socioeconomic-status differences were greater for Level II than for Level I abilities and (*b*) the correlation between Levels I and II and the regression of Level I upon Level II were greater in upper- than in lower-socioeconomic-status populations. These hypotheses were borne out by the present data, consisting of Level I measures (digit-span memory) and Level II measures (Lorge-Thorndike Intelligence Tests, verbal and nonverbal) obtained on all white and black pupils in Grades 4–6 in one school district. The largest effects were attributable to differences between the white population and the low-socioeconomic-status-black group.

The present study tests Jensen's Level I–Level II theory of mental abilities in a total school population. The theory has been tested heretofore only with specially selected samples from the population.

The theory and related evidence have been presented in detail elsewhere (Jensen, 1968, 1969, pp. 109–117, 1970a, 1970b, 1973, pp. 193–293; Jensen & Rohwer, 1968). Briefly, the theory involves two types of mental abilities, Level I and Level II, and their interaction with population (socioeconomic status and/or race) differences. Level I ability consists of rote learning and primary memory; it is the capacity to register and retrieve information with fidelity and is characterized essentially by a relative lack of transformation, conceptual coding, or other mental manipulation intervening between information input and output. Level II ability, in contrast, is characterized by mental manipulation of inputs, conceptualization, reasoning, and problem-solving; it is essentially the general intelligence (*g*) factor common to most complex tests of mental ability and standard tests of intelligence. Level I abilities are best measured by rote-learning tasks: serial learning, repeated trials of free recall of a number of successively presented familiar uncategorized objects, pictures, or nouns, and tests of short-term memory, such as digit span. Level II ability is best measured by tests of general intelligence that have a high general intelligence loading and especially those of the nonverbal, fluid-intelligence, culture-fair variety.

An interesting point about Level I and Level II abilities is their interaction with socioeconomic status and race, as has been shown in the articles cited previously. The first studies showed mainly that in groups of children selected for low Level II ability (IQs of 60–80), the low-socioeconomic-status children (white or black, although socioeconomic status and race are confounded in some studies) obtain markedly higher scores on Level I tests (usually approaching children with average IQs of 90–110) than are obtained by the middle- or upper-socioeconomic-status children with

[1] This research was supported by the Office of Economic Opportunity under Contract OEO 2404 and by grants to the University of California by the Berkeley Unified School District and the Sterling Morton Charitable Trust.

The author is grateful for the cooperation of the Berkeley Unified School District and for the able assistance of Wade Egbert, who supervised the testing, and Carol Treanor and Barbara Cerny, who conducted the computer analyses of the data.

[2] Requests for reprints should be sent to Arthur R. Jensen, Institute of Human Learning, University of California, Berkeley, California 94720.

the same low IQs. On Level I tasks, middle-socioeconomic-status children with low IQs perform more in accord with their low IQ, while low-socioeconomic-status children perform more like children of average IQ. This finding suggests a lower correlation between Level I and Level II ability of low-socioeconomic-status than in middle-socioeconomic-status populations. Also, it means that, in general, groups differing in socioeconomic status should differ less in Level I ability than in Level II ability. Thus it was suggested that if Level I ability could be made more important in the educative process, there might be a chance of diminishing the present large differences in scholastic performance associated with socioeconomic status and racial group differences in Level II ability, which is known to correlate highly with scholastic achievement in the prevailing system of education.

The earlier studies were based on a 2×2 analysis of variance design: high (or middle) socioeconomic status versus low socioeconomic status and high IQ (100–120) versus low IQ (60–80) thus forming four groups. Typically there were equal numbers of subjects (20–40) in each of the four groups. The low-IQ groups were often selected from classes for the educable mentally retarded with average IQs slightly below 70 (since 75 is the cutoff for admission to educable mentally retarded classes in California public schools). Because of the difficulty of matching low- and high-socioeconomic-status groups for high IQ, the "high-IQ" groups were usually only slightly above average (i.e., IQ of 105–110). The socioeconomic-status difference in Level I (learning) ability for the low-IQ groups was always highly significant, but the low- and high-socioeconomic-status groups of high IQ (i.e., about an IQ of 105–110) usually did not differ significantly, although the crossover or disordinal type of interaction usually appeared.

The present Level I–Level II hypotheses can be stated in their simplest form as follows:

1. Social classes do not differ, on the average, in Level I ability, but differ on Level II ability. (Another way of stating this is that Level I ability is not correlated with socioeconomic status and Level II ability is positively correlated with socioeconomic status.)

2. The regression of Level I upon Level II ability is greater (i.e., steeper slope of the regression line) in upper- and middle-socioeconomic-status populations than in low-socioeconomic-status populations. A less general corollary of this is that the correlation between Level I and Level II is greater in upper- and middle-socioeconomic-status populations than in low-socioeconomic-status populations. It is less general because restriction of the range of talent can affect the size of the correlation coefficient; whereas the slope of the regression line remains the same even if the distribution on one or both of the variables is truncated and the variance is thereby reduced. The correlation is lowered, therefore, but the slope of the regression line remains unchanged. The slope of the regression line (of Level I on Level II), therefore, is a more stable and fundamental datum. Thus, a proper test of the hypothesis should involve testing the difference in the regression of Level I upon Level II in low- and middle-socioeconomic-status groups.

The regression of Level II upon Level I has not been a part of the theory and cannot be inferred from the theory unless certain assumptions are made, assumptions for which at this point there seems to be no real theoretical basis. The lines of regression of Level II upon Level I can be determined only if we assume a precise value of the correlations between Levels I and II in low- and middle-socioeconomic-status groups. The theory posits no precise values, for no specific value exists for the general case. The correlations are merely population parameters, which may vary according to the populations sampled and the method of classifying individuals by socioeconomic-status group. The theory only posits that the regression coefficient (i.e., slope) of Level I on Level II is *greater* (i.e., steeper slope) in middle-socioeconomic-status than in low-socioeconomic-status populations. The posited difference thus is *directional* rather than precisely quantitative. If the variances are assumed to be the same in the middle- and low-socioeconomic-status groups (an assumption that is independent of the theory),

then the correlation between Levels I and II will be greater in the middle- than in the low-socioeconomic-status group. Assuming this to be the case, then, the regression lines (of Level II upon Level I) should gradually converge. But the convergence would be very gradual, and assuming realistic values of the Level I–Level II correlations in the low- and middle-socioeconomic-status groups, there would be no point within ±3 sigmas of the mean of a normal distribution (which includes 99.74% of the population) of Level I scores at which low- and middle-socioeconomic-status groups matched on Level I ability would be equal in Level II ability or where the low-socioeconomic-status group would exceed the middle-socioeconomic-status group in Level II ability. If the means of the low- and middle-socioeconomic-status groups were assumed to differ by 1 sigma and if the Level I–Level II correlations were .6 and .4 in the middle- and low-socioeconomic-status groups, respectively, we would have to match subjects from the two socioeconomic-status groups for Level I scores at least 5 sigmas below the common Level I mean in order for the low-socioeconomic-status subjects, on average, to equal or exceed the middle-socioeconomic-status subjects on Level II ability. But any subjects who were 5 sigmas below the mean on Level I ability would be in the range of severe mental defect, at the imbecile or idiot level, where the deficit is more likely due to a major gene or chromosomal anomaly or to organic damage, rather than to the normal variations in the polygenic and environmental determinants of mental variation that operate in the bulk of the population. For most of the normal population, the regression lines of Level II upon Level I for the two socioeconomic-status groups would be practically parallel. Estimating the point of convergence at 5 sigmas below the mean assumes linearity of regression all the way down into the range of severest mental defect, and since the causal factors in that range are different than for the rest of the distribution, such an assumption is quite unwarranted. Within reasonable boundary conditions for the operation of the theory, the lines of regression of Level II upon Level I should be pictured as almost parallel, with such slight convergence that the lines would not come together within the range of abilities normally found in the public schools. These hypotheses are depicted graphically in Figure 1.

As can be seen, the angles between the regression lines are different for the lower- and middle-socioeconomic-status groups (angles l and m), being smaller for the middle socioeconomic status. (The cosine of the angle between the regression lines is equal to the correlation coefficient when sigma is the same on both variables or the scores are standardized with the same sigma.)

3. The third element of the theory concerns the hierarchical relationship between Level II and Level I ability. The development of Level II ability, as well as Level II performance itself, is seen as having some functional dependence upon Level I ability, but the reverse is not true. For example, initial learning of the information and cognitive skills involved in Level II performance may depend in part upon short-term memory and its consolidation, which are Level I processes. Thus an individual with superior Level I ability will in the long run show better Level II performance than will a person with the same genetic and environmental potential for Level II ability but with

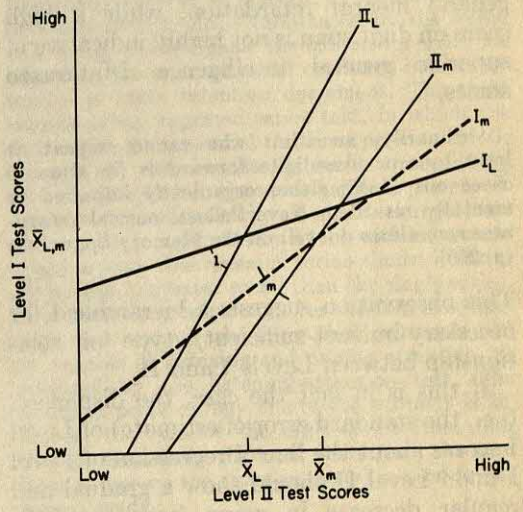


FIGURE 1. Hypothetical regression lines for relationship between Level I and Level II abilities in middle- and lower-class populations (angles l and m).

poorer Level I ability. Also, it seems reasonable to suppose that some short-term memory can be involved in solving Level II problems, such as Raven matrices items or the mental arithmetic subtest of the Wechsler scale, in which information must be retained in memory (i.e., Level I) while mental operations are being performed on it (i.e., Level II). A relatively pure Level I test, such as digit-span memory, on the other hand, can hardly be seen as depending upon the processes of abstraction, generalization, and conceptualization that are called for in Level II tests.

Another way of stating the hierarchical relationship between Levels I and II, that is, the functional dependence of Level II upon Level I ability, is to say that Level I is necessary but not sufficient for the development and operation of Level II ability. A consequence of this hierarchical formulation would be that one would seldom if ever find individuals with very high Level II ability who have very low Level I ability. The reverse, however, would not be uncommon, that is, persons with high Level I ability but low Level II ability. (In fact, quite extreme idiot savants of this type are known to exist.) As Matarazzo (1972, p. 204) has noted in connection with the clinical use of the Wechsler intelligence test, a low score on memory span for digits is highly related to general mental retardation, while a high score on digit span is not highly indicative of superior general intelligence. Matarazzo states,

> Ordinarily, an adult who cannot repeat at least four or five digits forward is [in about 9 cases out of 10] either organically impaired or mentally retarded. Nevertheless, mental retardates sometimes do well on the Memory Span Test [p. 205].

This observation suggests a hierarchical (or necessary-but-not-sufficient) type of relationship between Levels I and II.

If this is in fact the case, the dispersion (i.e., the standard error of estimate) of Level I scores about the line of regression of Level I upon Level II should show a gradual and regular decrease in going from lower to higher scores on Level II. Thus the relative magnitudes of the standard error of estimate of Level I scores for low and high scores on Level II provide a test of the hypothesis of hierarchical dependence of Level II upon Level I. That is, if the hypothesis is true, we should expect to find a larger dispersion of Level II scores in the lower range of Level I scores than in the higher range of Level I scores.

The purpose of the present study was to test each of these three main hypotheses (described under 1, 2, and 3) derived from the Level I–Level II theory of mental abilities.

Method

Subjects

The 2,612 subjects in this study consisted of virtually all the white (n = 1,489) and black (n = 1,123) children enrolled in regular classes of the fourth, fifth, and sixth grades from all 14 elementary schools of the Berkeley Unified School District in California.

The small percentage of children who were absent on the particular day that their class was tested are not included in this study. Also, test data on all children not classified in the school records (and according to their own parent(s)) as either white or black were excluded from the present study. (These excluded subjects, mostly Orientals, comprised about 10% of the total school population.)

The adult white population in this district is largely of middle or upper-middle socioeconomic status; the three largest employers (mostly of whites) are the university, the Lawrence Radiation Laboratory, and a large pharmaceutical firm, all of which employ workers with better than average education and socioeconomic status for the white population as a whole. The adult black population is predominantly lower-middle to low socioeconomic status, comprised largely of semiskilled and unskilled workers, although it is a somewhat higher-socioeconomic-status group than the black populations in the surrounding communities, with fewer unemployed and on welfare.

All tests were group administered to the regular classrooms by a staff of testers (3 whites and 3 blacks) who were specially employed and trained for this purpose. The white and black testers were assigned to classes at random. In any given class, the Level I and Level II tests were always administered by different testers on different days never more than one week apart. Thus the correlations between the Level I and Level II tests would not be systematically affected by any individual tester biases.

Tests

Control tests. Two different control tests were used, one in each of the two testing sessions. The

main purposes of the control tests were to set a good test-taking attitude in the class, emphasizing attention and effort while at the same time lessening tension and text anxiety by giving subjects tasks they could perform successfully simply by being attentive and trying their best.

The listening-attention test was given just before the Level I test (memory for numbers). The listening-attention test measures the child's ability to attend to and follow orally given directions paced at two-second intervals by means of a tape recording. The child is presented with an answer sheet containing 100 pairs of digits in sets of 10. The child listens to a tape recording that speaks one digit every two seconds. The child is required to put an X over the one digit in each pair that has been heard on the tape recorder. The purpose of this test is to determine the extent to which the child is able to pay attention to numbers spoken on a tape recorder, to keep his place in the test, and to make the appropriate responses to what he hears from moment to moment. Low scores on this test indicate that the subject is not up to validly taking the memory for numbers test, which follows immediately. High scores on the listen-attention test indicate that the subject has the prerequisite skills for taking the digit-span (memory for numbers) test. The listening-attention test thus is intended as a means for detecting students who, for whatever reason, are unable to hear and to respond to numbers read over a tape recorder. The test itself makes no demands on the child's memory, only on his ability for listening, paying attention, and responding appropriately—all prerequisites for the digit-memory test that follows.

The speed and persistence test (making Xs) was always given just before the Level II tests (Lorge-Thorndike IQ). The making Xs test is intended as an assessment of test-taking motivation. It gives an indication of the subject's willingness to comply with instructions in a group-testing situation and to mobilize effort in following those instructions for a brief period of time. The test involves no intellectual component, although for young children it probably involves some perceptual–motor skills component, as reflected in other studies by increasing mean scores as a function of age between Grades 1–5. Individual differences among children at any one grade level would seem to reflect mainly general motivation and test-taking attitudes in a group situation. Children who do very poorly on this test, it can be suspected, are likely not to put out their maximum effort on ability tests given in the same group situation, and to that extent, their ability test scores are not likely to reflect their real level of ability.

The making Xs test consists of two parts. On Part 1 the subject is asked simply to make Xs in a series of squares for a period of 90 seconds. In this part the instructions say nothing about speed. They merely instruct the child to make Xs. The maximum possible score on Part 1 is 150, since there are 150 squares provided in which the child can make Xs. After a two-minute rest period the child turns the page of the test booklet to Part 2. Here the child is instructed to show how much better he can perform than he did on Part 1 and to work as rapidly as possible. The child is again given 90 seconds to make as many Xs as he can in the 150 boxes provided. The gain in score from Part 1 to Part 2 reflects both a practice effect and an increase in motivation or effort as a result of the motivating instructions, that is, instructions to work as rapidly as possible.

Level I test. Previous studies have indicated that one of the most unambiguous and reliable Level I measures is digit-span memory. A specially devised test of such memory, which has much higher reliability than the short digit-span tests included in such general test batteries as the Stanford-Binet and the Wechsler, is the author's memory for numbers test. It has three parts. Each part consists of six series of digits going from four digits in a series up to nine digits in a series. The digit series are presented on a tape recording on which the digits are spoken clearly by a male voice at the rate of precisely one digit per second. The subjects write down as many digits as they can recall at the conclusion of each series, which is signaled by a bong. Each part of the test is preceded by a short practice test of three digit series in order to permit the tester to determine whether the child has understood the instructions, etc. The practice test also serves to familiarize subjects with the procedure of each of the subtests. The first subtest is labeled immediate recall. Here the subject is instructed to recall the series *immediately* after the last digit has been spoken on the tape recorder. The second subtest consists of delayed recall. Here the subject is instructed not to write down his response until 10 seconds have elapsed after the last digit has been spoken. The 10-second interval is marked by audible clicks of a metronome and is terminated by the sound of a bong, which signals the subject to write his response. The delayed recall condition invariably results in some retention decrement. The third subtest is the repeated series test, in which the digit series is repeated three times prior to recall; the subject then recalls the series immediately after the last digit in the series has been presented. Again, recall is signaled by a bong. Each repetition of the series is separated by a tone with a duration of one second. The repeated series almost invariably results in greater recall than the single series. This test is very culture fair for children in second grade and beyond, who know their numerals and are capable of listening and paying attention, as indicated by the listening-attention test. The maximum score on any one of the subtests is 39, that is, the sum of the digit series from four through nine. Only the total score (i.e., the sum of the scores on the three subtests) is used in the present study.

Level II tests. Level II was measured by the Lorge-Thorndike Intelligence Test (Level 3, Form B), which has two parts, verbal and nonverbal. This is a nationally standardized group-

administered test of general intelligence. In the normative population, which was intended to be representative of the nation's school population, the test has a mean IQ of 100 and a sigma of 16. The test is primarily a measure of reasoning ability; it has a high general intelligence saturation when factor analyzed with other mental ability tests, so it is deemed a good Level II test, especially the nonverbal part, which is based on pictorial problems and depends not at all upon reading skill or scholastic knowledge.

Results

Control Tests

On the listening-attention test there was no significant difference between the white and black groups in any grade. The mean number correct (out of 100) was above 98 for all groups in every grade, and the 25th, 50th, and 75th percentile score was 100 in each group at each grade. Since a perfect score on this test is 100, it is evident that the vast majority of subjects were motivated to do their best in the test situation and were capable of correctly hearing the numerals as spoken over the tape recording and of properly following directions and registering their responses on answer sheets. Practically all subjects obtained a perfect score. At this age level, there is no appreciable difference between the grades or between whites and blacks on the listening-attention test. Since the correlation between the listening-attention test and either the memory test or the Lorge-Thorndike Intelligence Tests is not significantly greater than zero in the white or black group, it is clear that no significant amount of the variance in these tests is attributable to differences in the kinds of sustained attentiveness and willingness to comply with instructions that are assessed by the listening-attention test.

On the speed and persistence test (making Xs), the black group scored significantly higher than the white group on both the first and second try and on the gain score (i.e., the difference between second try − first try), and these differences are fairly consistent across the three grades. These results, like those for the listening-attention test, indicate that at least equally good cooperation and effort in the test situation were put forth by the black subjects as by the white subjects. The lower quartile scores should be a most sensitive indicator of children who are not cooperating or putting out much effort, and at every grade the performance of black subjects equals or exceeds that of the white subjects. These results contradict the common notion that black children have a slower "personal tempo" or are less cooperative or more lackadaisical in a test situation. The correlation between making Xs and the memory for numbers and Lorge-Thorndike Intelligence Tests are close to zero in both racial groups.

Mean White–Black Difference in Memory (Level I) and Intelligence (Level II) Tests

The hypothesis in its most simple and extreme form states that low- and middle-socioeconomic-status groups differ in Level II but not in Level I ability. Table 1 shows the raw score means on the Level II and Level I tests in the white and black groups and (in the last column) shows the group difference in terms of the total within-groups variation. We see that although the white–black difference is highly significant both on the memory and on the intelligence tests, the difference on the intelligence tests is more than twice the difference on the memory test. It is thus unclear whether this finding disproves or supports the hypothesis. It would seem to disprove the "*no difference* on Level I" aspect of the hypothesis, and yet

TABLE 1

Raw Score Means and Standard Deviations on Intelligence (Level II) and Memory Tests (Level I) and Mean White–Black Differences in Sigma Units

Variable	White (n = 1,489)		Black (n = 1,123)		$(M_W - M_B)/\sigma$
	M	SD	M	SD	
Age (in months)	131.23	10.89	132.61	11.24	−.12
Intelligence					
Verbal	69.85	12.56	46.24	16.88	1.62
Nonverbal	63.12	10.83	43.47	14.50	1.57
Memory					
Immediate	23.33	6.41	18.75	6.61	.70
Repeat	26.89	5.81	23.40	6.56	.57
Delay	24.25	5.76	20.29	6.73	.64
Total	74.48	15.58	62.45	16.82	.75

Note. Sigma is the square root of the combined within-groups variance.

the results are consistent with the hypothesis in that the white–black difference on the memory test is very much less than the difference on the intelligence test. Since the theory also posits a correlation between Level I and Level II and a higher correlation in higher- than in lower-socioeconomic-status groups, we should expect a Level I difference between the present white and black groups if the IQ of the white group is above the white mean for the general population or if the present black group is below the general black mean IQ. In the general population, the groups differ by only about one sigma or 16 IQ points, while in our Berkeley population the difference is considerably greater. In terms of the Lorge-Thorndike national norms, the results of the present white group are verbal IQ, $M = 118.4$, $SD = 15.7$; nonverbal IQ, $M = 120.24$, $SD = 14.6$. The results of the present black group are verbal IQ, $M = 92.8$, $SD = 13.9$; nonverbal IQ, $M = 95.4$, $SD = 15.5$. The consequences of this difference between the groups used in the present study and the averages for the general United States population can be more easily discussed in the next section in connection with the regression of memory scores upon intelligence scores.

Regression of Memory upon Intelligence

The hypothesis predicts a steeper slope of the regression line of Level I (memory) scores upon Level II (intelligence) scores in the white group than in the black group. Figure 2 shows the relevant regression lines for the Lorge-Thorndike nonverbal scores. The graphical results are practically identical for the verbal scores.

A statistical test of departure from linearity was applied to all the regressions and none was found to depart significantly (at the .10 level) from linear regression. Though the linearity of the regression appeared to extend throughout the entire range of scores for both racial groups—a total range of more than 100 IQ points beginning at about an IQ of 50—the regression lines shown in Figures 2 and 3 were drawn to extend only over the range of scores that permits an unequivocal test of departure from linearity. (The *n*s at the very extremes of the distributions [less than the upper and lower 2.5%] are too small and scattered to permit statistical confidence of linearity in the regions beyond approximately the middle 95% of the distributions.)

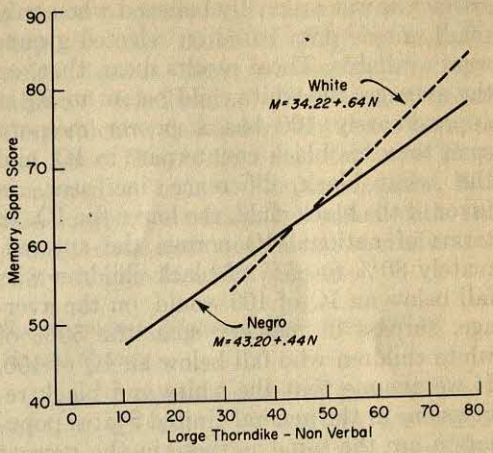


FIGURE 2. Regression of memory scores upon Lorge-Thorndike nonverbal intelligence raw scores in white and black groups.

We see in Figure 2 (and in verbal scores as well) that the regression is greater for whites than for blacks. For the Lorge-Thorndike verbal scores, the regression, b, is .58 for whites and .42 for blacks. A test[3] of the significance of the difference between the two slopes is highly significant ($t = 4.10$, $p < .001$). For the Lorge-Thorndike nonverbal scores, the white–black difference in slope is also highly significant ($t = 4.35$, $p < .001$).

According to the norms provided by the Lorge-Thorndike test manual, the point on the scale of raw scores at which the regression lines for the present white and black groups cross over is equivalent to a Lorge-Thorndike IQ of approximately 100, both for the verbal and the nonverbal tests. In the range of intelligence below an IQ of 100, the black children, on the average, surpass white children in memory scores; in the range above an IQ of 100, the white children surpass the black children in memory performance. The crossover in the above-average IQ range is clearly not a statistical

[3] The standard error of the regression (see Guilford, 1956) of Y on X, b_{yx}, is

$$\sigma_{b_{yx}} = \frac{\sigma_y}{\sigma_x}\sqrt{\frac{1 - r_{xy}^2}{N}}.$$

artifact as was originally believed when only small sample data based on selected groups were available. These results mean that, on the average, the white child below an IQ of approximately 100 has a poorer memory span than his black counterpart in IQ, and the white–black difference increases, in favor of the black child, the lower the IQ. In terms of national IQ norms, the approximately 80% to 85% of black children who fall below an IQ of 100 would, on the average, surpass in memory span the 50% of white children who fall below an IQ of 100. If we assume that the white and black regressions in the general United States population are the same as those in the present data and if the general white and black IQ means are 100 and 85, respectively, then, according to the regression equations in the present data, we should expect the white and black populations (which differ 1 sigma in IQ) to differ by only about .3 sigma to .4 sigma in memory span, in favor of whites. That is to say, the present data do not support the hypothesis of no white–black difference in Level I (here measured by the memory test), but the data do indicate a much smaller racial difference in memory than in IQ. This conclusion would, of course, not hold if the relative slopes of the regression lines for the two races are not about the same in the general population as in the Berkeley school population. The rather atypical nature of the Berkeley population with respect to mean Lorge-Thorndike IQ, especially in the white population, should make us wary of generalizing to the general population or to the populations of other communities with markedly different demographic and socioeconomic-status features than Berkeley.

Regression of Intelligence upon Memory

These regression lines present a very different picture from that of the regression of memory upon intelligence. As seen in Figure 3, the slopes of the regression lines for whites and blacks are parallel (the regression coefficients do not differ significantly), and they are separated by approximately 1.6 sigmas on the intelligence scales. (The results are virtually identical for the verbal scores.) Thus there is no point on the scale of memory

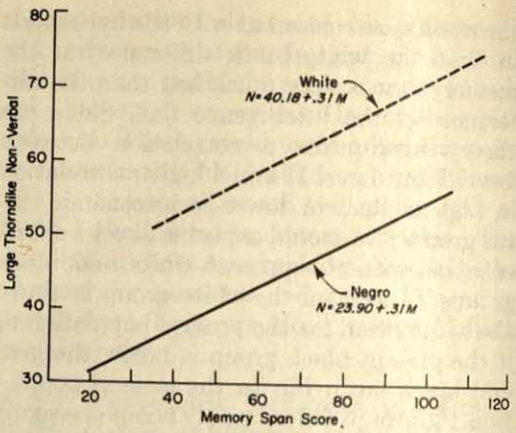


FIGURE 3. Regression of Lorge-Thorndike nonverbal raw scores upon memory scores in white and black groups.

scores at which equated groups of whites and blacks obtain equal intelligence scores. The picture is close to the hypothetical regression lines depicted in Figure 1. It would seem to be consistent with the hypothesis that Level I is necessary but not sufficient for the development and functioning of Level II. Why should white and black children with precisely the same memory performance differ by 1.6 sigmas on both the verbal and nonverbal intelligence measures? When matched for intelligence, on the other hand, whites and blacks are considerably more alike in memory, and they average just about the same in memory performance when matched on intelligence in the vicinity of an IQ of 100. In other words, it appears that if subjects have the intelligence, they have the memory; while if they have the memory, they do not necessarily have the intelligence.

Dispersion of Memory Ability as a Function of Intelligence

If it is true that intelligence depends upon memory but that the reverse does not hold, we should expect the dispersion of memory scores to show a systematic decrease going from low to high levels of intelligence. To test this hypothesis, the standard error of estimate of memory scores (i.e., the standard deviation of memory scores around the regression line of memory upon intelligence) was examined for systematic change over the full range of Lorge-Thorndike Intelligence Test scores, verbal and nonverbal. The

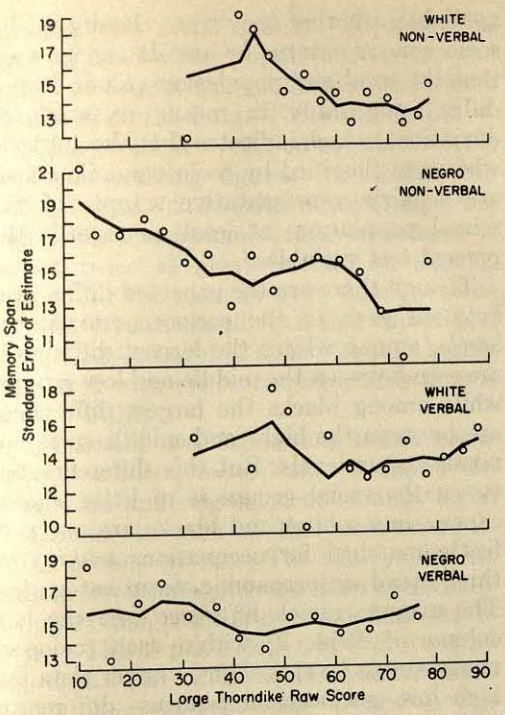


FIGURE 4. Memory score dispersion (standard error of estimate) as a function of Lorge-Thorndike verbal and nonverbal raw scores in white and black groups.

results are shown in Figure 4. Since the standard errors of estimates (indicated by circles) are rather erratic, their trend is better indicated by a moving average (the line going through the data points). For the nonverbal test the trend is clearly in accord with the hypothesis; that is, the standard error of estimate of the memory scores systematically decreases with increasing nonverbal intelligence. Bartlett's test for homogeneity of variances and a test of trend are both significant ($p < .01$) both for whites and for blacks. The results for the verbal test, however, yield only a faint suggestion of a decreasing standard error of estimate, and the trend is nonsignificant.

Thus the prediction based on the hypothesized hierarchical relationship between Level I and Level II is borne out by the nonverbal but not by the verbal test. Why should the two tests differ in this way? One can only speculate at this point. A possibility is that while both tests are highly saturated with general intelligence, the nonverbal test is more a measure of what Cattell (1971, chap. 5) calls "fluid" intelligence and the verbal test is more a measure of "crystalized" intelligence. The hypothesized hierarchical relationship between Level I and Level II may hold only for Level II as measured by tests of fluid intelligence. But this conjecture cannot be tested with the present data and must await a study specially designed for this purpose.

Socioeconomic-Status Differences within Racial Groups

A questionnaire sent to the home of every child in the study, as well as the school records, served as the basis for classifying subjects according to socioeconomic status. Among other items of family background information obtained from the parents was the current occupation (or last job held in case of the unemployed) of the head of the household. Since returns of the parental questionnaire were considerably less than 100%, especially in the black group, and not all the questionnaires that were returned had answered the occupation question, the sample size for the socioeconomic-status analysis was reduced and the remaining subjects cannot be regarded strictly as a random sample of the Berkeley school population because of the self-selection in answering the questionnaire. When parent's occupation was not given in the questionnaire, it was sought in school records, but often without success. If the parental occupation appeared in the school records, it almost invariably was given in the questionnaire and vice versa. Lack of information or ambiguity or doubt in the socioeconomic-status classification of a given occupation was cause for omitting subjects from the present analysis.

Parental occupations were first coded into 82 job description categories. These were then reduced to seven categories in terms of conventional socioeconomic-status rankings of the occupations. But in order to obtain large enough socioeconomic-status samples to allow tests of Level I–Level II correlations and regressions within each sample, these seven categories had to be reduced to three broad socioeconomic-status categories as follows:

High Socioeconomic Status
1. High-level administrators, supervisors, college teachers.
2. High-level professionals, engineers, physicians, etc.

Middle Socioeconomic Status
3. White collar occupations requiring college or technical training.
4. Self-employed, technicians, skilled craftsmen.
5. Merchants, managers of small business, service workers, contractors.

Low Socioeconomic Status
6. Manual workers.
7. Nonmanual workers, relatively unskilled, jobs ordinarily requiring less than a high school diploma.

The categories are admittedly crude and somewhat arbitrary, but would undoubtedly correlate highly with any of the various methods of socioeconomic-status classification.

Table 2 gives the means and standard deviations of the three socioeconomic-status groups within each race. The row labeled "total" is based on the subjects who were classifiable. The "population" row consists of all subjects on whom test data were available, whether they were classifiable by socioeconomic status or not. It can be seen that the total and population values do not differ appreciably in means or standard deviations, which indicates that the subjects who were classified by socioeconomic status are a fairly representative sample of the school population, at least as regards the present test variables.

Though there are the expected differences between each of the socioeconomic-status levels, among whites the largest differences are seen between the middle and low groups, while among blacks the largest differences are between the high- and middle-socioeconomic-status levels. But this difference between the racial groups is of little significance, since whites and blacks are not perfectly matched for occupations *within* the three broad socioeconomic-status categories. The average racial difference (see the last column of Table 2) within each socioeconomic-status level is slightly larger than the high–low socioeconomic-status differences within each race for the Lorge-Thorndike verbal and nonverbal scores. For the memory scores, on the other hand, the high–low socioeconomic-status difference within each racial group is greater than the difference

TABLE 2

MEANS AND STANDARD DEVIATIONS OF INTELLIGENCE AND MEMORY RAW SCORES OF SOCIOECONOMIC-STATUS LEVELS WITHIN RACIAL GROUPS AND MEAN DIFFERENCE IN SIGMA UNITS

Test	Socioeconomic status	White			Black			$(M_W - M_B)/\sigma$
		n	*M*	*SD*	*n*	*M*	*SD*	
Lorge-Thorndike verbal	High	763	71.6	10.40	38	58.4	13.22	1.25
	Middle	287	70.9	9.98	43	50.5	15.88	1.87
	Low	215	60.3	16.42	414	45.7	15.99	.90
	Total[a]	1,265	69.6	12.24	495	47.2	16.19	1.66
	Population[b]	1,489	69.9	12.56	1,123	46.2	16.88	1.63
Lorge-Thorndike nonverbal	High	763	65.3	8.75	38	52.3	11.36	1.46
	Middle	287	63.6	9.61	43	46.6	12.04	1.71
	Low	215	55.6	14.82	414	43.0	13.63	.90
	Total	1,265	63.4	10.72	495	44.1	13.59	1.66
	Population	1,489	63.1	10.83	1,123	43.5	14.50	1.56
Memory (total score)	High	763	74.7	15.23	38	69.0	16.12	.37
	Middle	287	73.7	14.01	43	63.4	16.65	.72
	Low	215	65.5	15.28	414	60.9	16.76	.28
	Total	1,265	72.9	15.34	495	61.7	16.86	.71
	Population	1,489	74.5	15.58	1,123	62.5	16.82	.74

Note. $M_W - M_B/\sigma$, where σ is the square root of the combined within-groups variance.
[a] Total of all subjects who were classified by socioeconomic status.
[b] The entire school population in Grades 4–6.

between the racial groups. Expressed in units of the white population sigma, the high–low socioeconomic-status difference for whites on the Lorge-Thorndike verbal = .90, on nonverbal = .90, on memory = .59; the corresponding figures for blacks are 1.02, .86, and .52. In both racial groups, the high–low socioeconomic-status difference is almost twice as great for the intelligence tests as for the memory test, which accords with the hypothesis at least in a directional sense, that is, Level II ability is more highly correlated with socioeconomic status than is Level I ability.

Correlations and regressions within socioeconomic-status groups. Table 3 shows the correlation and regression of memory upon intelligence in each of the socioeconomic-status groups by race. The theory predicts higher correlations and regression coefficients in upper- than in lower-socioeconomic-status groups. This is not completely borne out by the data. The white-socioeconomic-status groups show no systematic trends in this respect, but the black-socioeconomic-status groups show the predicted trend, that is, lower correlations and regressions with lower socioeconomic status. The black-high- and middle-socioeconomic-status groups both appear quite different from the black-low-socioeconomic-status group. The differences

TABLE 3

CORRELATIONS (r) AND REGRESSION COEFFICIENTS (b) OF MEMORY SCORES UPON LORGE-THORNDIKE VERBAL AND NONVERBAL RAW SCORES IN SOCIOECONOMIC-STATUS LEVELS WITHIN RACIAL GROUPS

Socioeconomic status	White				Black			
	Verbal		Nonverbal		Verbal		Nonverbal	
	r	b	r	b	r	b	r	b
High	.376	.551	.361	.628	.552	.673	.594	.843
Middle	.414	.581	.324	.472	.408	.428	.496	.686
Low	.591	.550	.605	.624	.391	.410	.311	.382
Total[a]	.464	.582	.442	.632	.419	.436	.365	.453
Population[b]	.466	.578	.443	.637	.420	.419	.372	.436

[a] Total of all subjects who were classified by socioeconomic status (white, n = 1,265; black, n = 495).

[b] The entire school population in Grades 4–6 (white, n = 1,489; black, n = 1,123).

of regression coefficients between all 15 possible contrasts of Socioeconomic Status × Race groups in Table 3 were subjected to t tests (two-tailed) to determine their significance. Only three of the contrasts are significant beyond the .05 level (two-tailed), all involving only the nonverbal test: high-socioeconomic-status white – low-socioeconomic-status black ($p < .01$); low-socioeconomic-status white – low-socioeconomic-status black ($p < .01$); high-socioeconomic-status black – low-socioeconomic-status black ($p < .03$).

All the significant differences involve exclusively the low-socioeconomic-status-black group, and the only significant within-race socioeconomic-status difference is between high- and low-socioeconomic-status blacks. The difference in regressions, therefore, appears to involve race more than socioeconomic status, or a combination of race and socioeconomic-status effects, since the low-socioeconomic-status-black group is undoubtedly somewhat below the low-socioeconomic-status-white group in socioeconomic status. The regressions of the high-and middle-socioeconomic-status-black groups do not differ significantly from those of the white groups.

DISCUSSION

The present study examined three main aspects of the Level I–Level II theory of mental ability, namely (*a*) the relative magnitudes of socioeconomic status and white–black differences in Level I and Level II abilities, (*b*) socioeconomic status and racial differences in the correlation between Levels I and II and in the regression of Level I upon Level II, and (*c*) the hierarchical (i.e., necessary-but-not-sufficient) functional dependence of Level II performance on Level I ability.

The *a* theory as originally stated in its simplest form predicts a socioeconomic-status difference in Level II ability but not in Level I ability. This formulation, however, was intended more as an unambiguous basis for a directional prediction than as a precise expectation of reality, for in reality it is, of course, most improbable that there is "no difference" between any two populations in any given trait. So the realistic issue is the

relative magnitudes of differences between populations in Levels I and II. In accordance with previous findings, it was found that the white and black groups, and to a slightly lesser degree the high- and low-socioeconomic-status groups within each race, differed much more, on the average, in Level II than in Level I ability. The exact size of the differences, of course, depends upon the particular populations being compared and is not regarded as an intrinsic aspect of the theory, the main point of which is that populations can differ in these two classes of ability and that the direction of the difference in socioeconomically stratified populations is such that the higher and lower groups will show a greater difference on Level II than on Level I. The reason for this, according to the theory, is that social mobility in an industrialized society is more dependent upon Level II than upon Level I abilities. In the present study the white–black differences are larger than the socioeconomic-status differences within the racial groups, but the point is ambiguous here since the average socioeconomic-status difference between the races is probably greater than the high–low socioeconomic-status differences within each racial group. The strict criteria for socioeconomic-status classification used here resulted in the inclusion of a peculiarly small percentage of the black population in the high- and middle-socioeconomic-status categories. It would be advisable in future studies to have socioeconomic-status ratings on a continuous scale based on a large number of home background factors, which might reflect more closely the nature of the child's environment than does merely the occupational classification of his parents.

The hypothesized higher correlation of *b* between Levels I and II in the white than in the black group was fully borne out by the data, as was also the predicted higher regression of Level II upon Level I. The effect is largely attributable to the difference between the entire white sample and the low-socioeconomic-status-black group, which constituted the vast majority of the present black sample. The high-socioeconomic-status- and middle-socioeconomic-status-black groups do not differ significantly from the white population in this respect but differ significantly from the low-socioeconomic-status-black group.

The cause of different Level I–Level II correlations (or regressions) in different populations has not yet been established and at present can only be hypothesized. There are several possible causes of correlation and they are not mutually exclusive: (*a*) part–whole functional dependence, that is, one behavior may be a subunit of some other behavior, such as shifting gears smoothly and passing a driver's test consisting of driving in traffic with an examiner present; (*b*) hierarchical functional dependence, that is, one behavior is prerequisite to another or one is functionally dependent upon another, as skill in working problems in long division is dependent upon skill in multiplication; (*c*) environmental correlation between the behaviors, that is, cultural contingencies may be such that when one behavior is learned another is also likely to be learned, even though there is no functional connection between the two behaviors, for example, a knowledge of baseball and a knowledge of football; and (*d*) genetic correlation between behaviors due to common assortment of their genetic underpinnings through selection and homogamy and pleiotropism (one gene having two or more phenotypic effects). The rather low degree of correlation between our Level I and Level II tests suggests that there is little functional dependence, and this could be proved conclusively if one could find a group of subjects that reliably showed a zero correlation between Level I and Level II. The fact of quite large and significant differences in Level I–Level II correlations in various populations is also inconsistent with wholly functional or part–whole dependence as a cause of the correlation. Some substantial part of the correlation therefore, must be attributable to other causes. If the cause is common environmental influences on the Level I and Level II tests, it is hard to imagine what these influences might be and why, if they are common, there should be such large group mean differences in Level II ability and not in Level I. The most reasonable hypothesis at this point would seem to be that the correlation is due only slightly to functional dependence of

Level II upon Level I and mostly to a common genetic assortment on both factors, that is, a genetic correlation in the population between two broad classes of ability with different genetic underpinnings. If this were the case, we might find a wide range of correlations in different populations; one conceivably might even find a group in which the correlation is negative. This would tend to rule out pleiotropy and would suggest independent mechanisms under independent genetic control underlying Level I and Level II. Specially designed studies would be required to test such a hypothesis.

The test of the *c* hypothesis of hierarchical dependence of Level II upon Level I yielded significant evidence consistent with the hypothesis in the case of the nonverbal intelligence test but not the verbal. In any case, there does not appear to be evidence of any strong degree of functional dependence between the abilities; quite low or high scores on the one ability are not incompatible with a high or low score on the other, though there is a tendency for low intelligence – high memory to be more frequent than the opposite combination of abilities, especially for nonverbal intelligence.

In the present study, Level I ability was measured by three slightly differing forms of a single type of test—digit-span memory. In other studies different tests have been used—paired-associate learning, serial learning, and free recall of pictures and objects—all with similar results generally consistent with the formulation of the two-level theory.

REFERENCES

CATTELL, R. B. *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin, 1971.

GUILFORD, J. P. *Fundamental statistics in psychology and education.* (3rd ed.) New York: McGraw-Hill, 1956.

JENSEN, A. R. Patterns of mental ability and socioeconomic status. *Proceedings of the National Academy of Sciences of the United States of America,* 1968, **60,** 1330–1337.

JENSEN, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review,* 1969, **39,** 1–123.

JENSEN, A. R. Hierarchical theories of mental ability. In B. Dockrell (Ed.), *On intelligence.* Toronto, Canada: Ontario Institute for Studies in Education, 1970. (a)

JENSEN, A. R. A theory of primary and secondary familial mental retardation. In N. R. Ellis (Ed.), *International review of research in mental retardation.* Vol. 4. New York: Academic Press, 1970. (b)

JENSEN, A. R. *Genetics and education.* New York: Harper & Row, 1973.

JENSEN, A. R., & ROHWER, W. D., JR. Mental retardation, mental age, and learning rate. *Journal of Educational Psychology,* 1968, **59,** 402–403.

MATARAZZO, J. D. *Wechsler's measurement and appraisal of adult intelligence.* (5th ed.) Baltimore, Md.: Williams & Wilkins, 1972.

(Received January 29, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 112-116

STIMULUS AND RESPONSE VARIABLES IN CHILDREN'S LEARNING OF GRAPHEME–PHONEME CORRESPONDENCES

GEORGE MARSH,[1] AND PETER DESBERG

California State College, Dominguez Hills

LAWRENCE KENT FARWELL

University of California, Santa Barbara

The effects of stimulus and response abstractness and similarity were investigated in children's paired-associate learning of grapheme–phoneme correspondences. In the first experiment, stimulus and response abstractness were investigated in a factorial design. The results indicated that the major source of learning difficulty was response abstractness. In a second study, stimulus and response similarity were also varied factorially. Neither factor had a significant effect on the number of correct responses. The implications of the results for teaching grapheme–phoneme correspondences were discussed.

The purpose of the present study was to investigate the effects of stimulus and response abstractness and similarity on children's learning of grapheme–phoneme correspondences. Both of these factors are known to be an important factor in adult paired-associate learning, but little research on these variables has been conducted with children. Keppel (1964), in his review of children's verbal learning, commented upon the absence of studies investigating these factors with children.

Since Keppel's (1964) article, studies by Paivio and Yuille (1966) and Dilley and Paivio (1968) have investigated the effects of meaningfulness and abstractness on children's paired-associate learning. However, no study has investigated the effects of these variables in the context of tasks directly related to beginning reading.

Although young children perform relatively well on paired-associate learning of concrete materials such as picture–picture pairs (e.g., Jensen and Rowher, 1965), they have considerable difficulty in learning grapheme–phoneme correspondences that from their point of view are relatively abstract and meaningless (Calfee, Chapman, & Venezky, 1972; Marsh & Sherman, 1969).

Feldman, Johnson, and Mast (1972) have recently reported that performance on concrete paired-associate tasks is unrelated to school achievement, while performance on abstract tasks shows a substantial correlation with school achievement. This would be expected since most paired-associate tasks in school are relatively abstract.

Experiment 1

The purpose of the first study was to determine whether children's relatively poor performance on grapheme–phoneme correspondence tasks is primarily a function of stimulus or response concreteness. A factorial design varying stimulus and response concreteness which was similar to that used by Cieutat, Stockwell, and Noble (1958) with adults and Dilley and Paivio (1968) with children was employed. This involves four combinations of stimulus and response concreteness: high–high, high–low, low–high, and low–low. In the high–high group, the stimulus and response items were both

[1] Requests for reprints should be sent to George Marsh, California State College, 1000 East Victoria Street, Dominguez Hills, California 90246.

familiar pictures (Group Picture–Picture); in the high–low group, the stimulus was a picture and the response was a letter sound (Group Picture–Letter); in the low–high group, the stimulus was a letter and the response item was a picture (Group Letter–Picture); and in the low–low group, the stimulus was a letter and the response its appropriate letter sound (Group Letter–Letter).

Method

Subjects. The subjects were 40 nursery school children between the ages of 53 and 65 months with a mean age of 61 months. Each child was pretested for knowledge of grapheme–phoneme pairs. Three subjects were excluded from the experiment on this basis.

Stimulus materials. The stimulus materials for the abstract conditions were 12 lowercase graphemes approximately 2 inches high printed on 4 × 6 inch cards. The graphemes consisted of 6 stop consonants (p, t, k, b, d, and g) and 6 continuant consonants (f, v, s, n, z, and m). The responses in this condition consisted of the appropriate phoneme associated with each grapheme. The stop phonemes were followed by a reduced schwa (e.g., /pə/) since they are impossible to produce in complete isolation. The stimulus materials for the concrete conditions consisted of 12 pictures of common objects taken from a children's alphabet book (telephone, bird, lion, apple, keys, flowers, shoes, doll, clock, house, rabbit, and balloons).

Procedure. The subjects were randomly assigned to one of the conditions (picture–picture, picture–letter, letter–picture, or letter–letter) with restriction of equal *n*s in each group ($n = 10$). When the subject was seated across from the experimenter, he was pretested for knowledge of grapheme–phoneme pairs and then given the following instructions.

> We're going to play a game. First I'll tell you what goes with each of these pictures (letters). Then when I show you the picture (letter) again, you tell me what goes with it. After you answer I'll tell you what goes with the picture (letter).

The experimenter then selected a list of 6 items previously made up from the pool of materials available. The type of list constructed depended on the group in which the subject was assigned. In the picture–picture condition, the 12 pictures were randomly paired to form the 6-item list; in the picture–letter condition, 6 pictures were randomly paired with 6 of the 12 phonemes; in the letter–picture condition, 6 graphemes were randomly paired with 6 pictures; and in the letter–letter condition, 6 of the 12 grapheme–phoneme pairs were selected randomly.

The subjects were then given a familiarization trial (during which the experimenter pronounced the phonemes and the picture name responses associated with each stimulus). Following the familiarization trial, each subject was given 10 training trials by a method of anticipation in which the child was given approximately 10 seconds to respond. After a response or the anticipation period had elapsed, the experimenter presented the correct stimulus–response pair. The responses were recorded as either correct, omission, or intrusion.

Results

The mean number of correct responses, omission errors, and intralist intrusion errors are shown in Table 1.

The correct responses were analyzed in a 2 × 2 × 10 factorial analysis of variance. The between-subjects factors were stimulus and response abstractness and the within-subjects factor was trials. The effect of stimulus abstractness was not significant ($F < 1$). The effect of response abstractness was significant ($F = 41.49$, $df = 1/36$, $p < .001$).

The trials effect was significant ($F = 32.50$, $df = 9/324$, $p < .001$). The interaction between trials and stimulus abstractness was not significant ($F = 1.19$, $df = 9/324$) but the interaction between type of response and trials was significant ($F = 3.98$, $df = 9/324$, $p < .01$).

The omission errors were also analyzed in a 2 × 2 factorial analysis of variance. Again, the effect of the stimulus factor was not significant ($F < 1$). The effect of re-

TABLE 1

MEAN NUMBER OF CORRECT RESPONSES, OMISSION ERRORS, AND INTRALIST INTRUSION ERRORS

Group	Correct responses	Omission errors	Intrusion errors
Experiment 1			
Letters–letter sounds	20.6	27.7	11.7
Pictures–letter sounds	16.2	31.9	11.9
Letters–picture names	41.0	8.3	10.7
Pictures–picture names	46.9	7.9	5.2
Experiment 2			
Low visual–low acoustic	17.5	13.7	8.8
High visual–low acoustic	15.2	15.5	9.3
Low visual–high acoustic	16.0	12.8	11.2
High visual–high acoustic	12.7	12.9	14.4

sponse factor was significant ($F = 42.8$, $df = 1/36$, $p < .001$). The interaction between stimulus and response factors was not significant ($F < 1$).

A similar 2 × 2 factorial analysis of variance was carried out on the intrusion errors. Neither the main effects nor the interaction was significant.

Experiment 2

The purpose of the second study was to investigate the effect of intralist stimulus and response similarity on children's learning of grapheme–phoneme correspondences. The design of the study was similar to that of the first experiment except that stimulus and response similarity, rather than abstractness were combined factorially to form four conditions: high visual and high acoustic similarity (Group High Visual–High Acoustic); high visual and low acoustic similarity (Group High Visual–Low Acoustic); low visual and high acoustic similarity (Group Low Visual–High Acoustic); and low visual and low acoustic similarity (Group Low Visual–Low Acoustic).

Method

Subjects. The subjects were 40 nursery school children between the ages of 51 and 64 months with a mean age of 58 months. As in the first experiment, the subjects were screened for their knowledge of the grapheme–phoneme correspondences used in the study. Two children were excluded from the study on this basis.

Stimulus materials. The paired-associate lists were four-item grapheme–phoneme pairs that were selected on the basis of graphemic similarity using the data of Dunn-Rankin (1968) to scale visual similarity and the data on isolated phoneme discriminability collected by Marsh and Sherman (1971) to scale acoustic similarity. The lists were constructed so that each grapheme would be visually similar to one other grapheme in the high-visual-similarity list and each phoneme would be acoustically similar to one other phoneme in the high-acoustic-similarity list (See Table 2).

The stimuli were lowercase letters approximately 2 inches high printed on 4 × 6 inch cards. The responses consisted of the pronunciation of the letter sounds (phonemes). The phonemic response to the grapheme *t* in Group High Visual–Low Acoustic and to the *h* in Group Low Visual–High Acoustic was the fricative /θ/ which is spelled with a th and is acoustically confusable with /f/.

Procedure. The procedures for assigning subjects to groups, pretesting for knowledge of letter sound associations, and training by the method of anticipation was identical to procedures followed in the first experiment.

Results

The mean number of correct responses, omission errors, and intralist intrusion errors for the four groups is shown in Table 1. A 2 × 2 × 10 analysis of variance showed that neither stimulus similarity ($F = 1.76$, $df = 1/36$, $p < .01$) nor response similarity ($F = 1.04$, $df = 1/36$, $p < .01$) was significant. The trials effect was significant ($F = 19.69$, $df = 9/324$, $p < .001$), but there was no significant interaction between trials and the stimulus and response variables.

A 2 × 2 factorial analysis of variance on the omission errors showed no significant main effects or interactions.

A 2 × 2 factorial analysis of variance on the intrusion errors showed that the main effect for stimulus similarity was not significant ($F = 1.50$, $df = 1/36$), while the main effect for response similarity was significant ($F = 6.36$, $df = 1/36$, $p < .05$). The interaction between stimulus and response similarity was not significant ($F < 1$).

Discussion

The results of Experiment 1 indicate that the major factor in the children's relatively

TABLE 2

Paired-Associate Lists and Responses for the Visual–Acoustic Similarity Groups

Similarity group	List	Response
High visual–high acoustic	m, n, f, t	/m/, /n/, /f/, /θ/
High visual–low acoustic	d, b, m, w	/də/, /bə/, /m/, /w/
Low visual–high acoustic	s, f, v, h	/s/, /f/, /v/, /θ/
Low visual–low acoustic	b, s, m, k	/bə/, /s/, /m/, /k/

poor performance on grapheme–phoneme correspondence pairs as compared with concrete tasks such as picture–picture pairs is response availability. It made no significant difference if the stimuli were graphemes or familiar pictures, but it made a large difference in performance if the responses were phonemes as opposed to picture names. This conclusion is supported by both the analysis of correct responses and the analysis of the omission errors. These results are in agreement with the results of most adult studies of paired-associate learning. As Underwood and Schultz (1960) point out, the subject does not have to produce the stimulus but does have to produce the response; thus implicating response availability as the major factor in paired-associate task difficulty. However, the present results are in conflict with a previous study by Dilley and Paivio (1968) which found that pictures as stimuli significantly improve performance as compared to spoken words as stimuli, while pictures as response items retarded learning as compared to spoken words as responses. The difference in the outcome of this study and that of Dilley and Paivio (1968) is most probably a function of the materials used. While pictures as stimuli may mediate association to spoken words, it is hard to see how they could have this effect when responses are phonemes. Spoken words as responses are apparently more effective than pictures as response items because the subject was required to produce a picture name (i.e., a spoken word) rather than an image of the picture in the Dilley and Paivio experiment.

Many research programs on beginning reading concentrate on the visual aspects of the reading task. This emphasis appears to be misdirected, at least when dealing with simple grapheme–phoneme correspondences. The major practical implication of the present research is that efforts to facilitate the acquisition of grapheme–phoneme correspondences by prereading children should be directed toward increasing response availability rather than stimulus differentiation or familiarity. Prereading children apparently have little difficulty in dealing with letters as stimuli but a great deal of difficulty in dealing with letter sounds as responses.

In Experiment 2, although both the graphemes and phonemes were selected to maximize the effect of similarity, there was no overall effect of these variables on the number of correct responses. Underwood, Runquist, and Schultz (1959) have discussed the dual and opposed effects of response similarity on learning. On the one hand, increases in response similarity may facilitate learning by increasing response availability, while at the same time interfering with the associative component due to intrusion errors. In the present experiment, there was no significant effect of similarity on response availability as indexed by omission errors; however, there was a small but statistically significant negative effect of response similarity on the associative component as indexed by the number of intrusion errors.

According to Underwood (1961, p. 215), an analogous effect could occur when stimulus similarity is manipulated in pairs of stimuli, as has been done in several experiments including the present one. Because of the constraints imposed by natural language materials (grapheme–phoneme pairs), it was only possible in the present experiment to manipulate similarity between pairs of items in the list rather than across the list as a whole. Underwood (1961, p. 215) concludes that all stimulus items in the list should be similar to each other in order to obtain a stimulus similarity effect.

From a practical point of view, it appears that including two pairs of high-similarity grapheme–phoneme pairs in the same instructional block will have no significantly detrimental effect on learning.

REFERENCES

Calfee, R., Chapman, R., & Venezky, R. How a child needs to think to learn to read. In L. Gregg (Ed.), *Cognition in learning and memory*. New York: Wiley, 1972.

Cieutat, V. J., Stockwell, F. E., & Noble, C. E. The interaction of ability and amount of practice with stimulus and response meaningfulness (m, m') in paired-associate learning. *Journal of Experimental Psychology*, 1958, **56**, 193–202.

Dilley, M., & Paivio, A. Pictures and words as stimulus and response items in paired-associate

learning of young children. *Journal of Experimental Child Psychology*, 1968, **6**, 231–240.

DUNN-RANKIN, P. The similarity of lower-case letters of the English alphabet. *Journal of Verbal Learning and Verbal Behavior*, 1968, **7**, 990–995.

FELDMAN, D. H., JOHNSON, L. E., & MAST, T. A. Correlation of paired-associate performance with school achievement as a function of task abstractness. *Journal of Educational Psychology*, 1972, **63**, 423–428.

JENSEN, A. R., & ROWHER, W. D. Syntactical mediation of serial and paired-associate learning as a function of age. *Child Development*, 1965, **36**, 601–608.

KEPPEL, G. Verbal learning in children. *Psychological Bulletin*, 1964, **61**, 63–80.

MARSH, G., & SHERMAN, M. *Transfer from words to word components and vice versa.* (Tech. Rep. 23) Inglewood, Calif.: Southwest Regional Laboratory, 1969.

MARSH, G., & SHERMAN, M. *Children's discrimination and production of phonemes in isolation and in words.* (Technical Memorandum 2-71-07) Inglewood, Calif.: Southwest Regional Laboratory, 1971.

PAIVIO, A., & YUILLE, J. C. Word abstractness and meaningfulness and paired-associate learning in children. *Journal of Experimental Child Psychology*, 1966, **4**, 81–89.

UNDERWOOD, B. J. An evaluation of the Gibson theory of verbal learning. In C. N. Cofer and B. S. Musgrave (Eds.), *Verbal learning and verbal behavior*. New York: McGraw-Hill, 1961.

UNDERWOOD, B. J., RUNQUIST, W. N., & SCHULTZ, R. W. Response learning in paired-associate lists as a function of intralist similarity. *Journal of Experimental Psychology*, 1959, **58**, 70–78.

UNDERWOOD, B. J., & SCHULTZ, R. W. *Meaningfulness and verbal learning*. Philadelphia: Lippincott, 1960.

(Received February 5, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 117–122

BIASING EFFECTS OF DIAGNOSTIC LABELS AND SEX OF PUPIL ON TEACHERS' VIEWS OF PUPILS' MENTAL HEALTH[1]

PHYLLIS F. HERSON[2]

University of Maryland

One hundred and eighty teachers were asked to respond to a Thurstone-type scale measuring their view of the degree of psychological incapacitation manifested by four hypothetical pupil cases. Variations in methods of describing pupils assigned randomly to the subjects were (*a*) diagnostic labels only, (*b*) behavioral descriptions only, and (*c*) behavioral descriptions with diagnostic labels affixed. In addition, designations in half of the cases in each variation were changed from masculine to feminine. A $4 \times 3 \times 2$ analysis for repeated measures indicated that when diagnostic labels were used, mean scores on the incapacitation scale were significantly higher than when descriptions alone were used ($p < .005$). No significant overall difference in mean scores was found between the two pupil sexes.

Currently used mental health terminology has come to have great social, moral, judicial, and even political implications in our society, directly affecting decisions in such areas as hospitalization, imprisonment, and educational and vocational advancement and more subtly influencing a wide array of social interactions. The focus of this study is upon the implications of the use of such terminology in a school setting.

While the importance of the classification or labeling process has been given recognition in several theoretical propositions, the most formalized one being that of Scheff (1966), empirical support for this phenomenon must be derived from peripheral areas such as psychological studies of expectancy effects in laboratory experiments (for example, see Cordaro & Ison, 1963). Few empirical studies were found that attempted to assess the effects of the use of mental health type labels in such natural settings as the school. The effects of such labels upon the attitudes and policies of teachers, however, would seem to be of particular concern because of the relative youth and malleability of the people with whom they interact on a day-to-day basis. Teachers, furthermore, are in a crucially pivotal position with regard to labeling in that they are both consumers of and, in some instances, originators of labels through the media of cumulative records, case conferences, and informal communications with other staff members.

Some evidence, most notably from the Rosenthal and Jacobson study (1968), was found to suggest that labels denoting academic potential may become self-fulfilling prophecies because of subtle changes, which result from the use of labels, in teachers' interactions with pupils. No study was found, however, that attempted to assess such changes in teachers' views of pupils as a result of the use of mental health type labels.

An examination of the literature dealing with the semantics of mental health terminology, however, reveals that such labels may carry negative or misleading implications. The replacement of the demonological model of mental illness with the medical model has, it would seem, not fully eradi-

[1] This article is based on the author's dissertation submitted in partial fulfillment of the requirements for the doctoral degree at the University of Maryland.

[2] Requests for reprints may be sent to Phyllis F. Herson, 9104 Walden Road, Silver Spring, Maryland 20901.

cated older associations of mental illness with wild, bizarre behavior (Nunnally, 1961) and has, in addition, brought about some misleading implications of its own. Among these are a view of the course of mental illness as a rigidly predetermined process and an emphasis upon causal agents within the individual as opposed to those in the social environment (Sarbin, 1967; Szasz, 1960). Currently used diagnostic labels, it has been further pointed out, are too broad, which results in a tendency to attribute all of the behavior associated with a category to an individual, even though only part of the behavior led to the original labeling (Ullmann & Krasner, 1969).

It was further established in the literature that teachers today, to a greater extent than was formerly true, tend to be sensitive and aware of mental health phenomena and to look to mental health professionals as authorities and as sources of help (Bentz, Edgerton, & Miller, 1969). Though this is generally regarded with favor as indicative of greater enlightenment than was formerly the case, nevertheless, it is suggested that this heightened sensitivity and awareness could have the effect of rendering teachers more highly suggestible with regard to mental health labels than were their seemingly more skeptical forebears.

The literature further suggests that the effects of verbal expectancy statements or labels would seem to be mitigated by a variety of observer characteristics such as sex (Stukat, 1958) and degree of training, experience, or familiarity with the phenomenon under observation (Ingraham & Harrington, 1966). In addition, evidence suggests that certain pupil characteristics such as sex may influence teachers' attitudes and policies toward deviant behavior (Yamamoto & Dizney, 1967).

The purposes of this study then were (*a*) to determine whether the use of mental health type diagnostic labels has a biasing effect upon teachers' assessments of the degree of incapacitation of hypothetical pupils, (*b*) to determine whether teachers' assessments of the degree of incapacitation differ as a function of the sex of the pupil, and (*c*) to determine whether sex of the teacher or length of teaching experience are mediating variables with regard to these effects.

Method

Subjects

Subjects were 64 male and 116 female elementary and secondary school teachers enrolled in summer classes at the University of Maryland. Length of teaching experience ranged 0–36 years.

Procedure

From the population of 180, samples of 30 were randomly assigned to six treatment groups or six variations in stimulus material. Each respondent was presented with descriptions of four types of hypothetical pupils or cases—marginally retarded, depressed neurotic, emotionally disturbed, and paranoid schizophrenic—in one of six treatment formats. Pupils or cases were described by means of (*a*) diagnostic labels only, (*b*) behavioral descriptions only, and (*c*) behavioral descriptions with diagnostic labels affixed. Each of these methods of description was further modified by changing appropriate designations and pronouns from masculine to feminine. For each of the four cases, subjects were asked to respond to a Thurstone-type scale developed by the investigator to measure their view of the degree of psychological incapacitation manifested by the hypothetical pupil.

Materials

Case descriptions and labels. The development and validation of the four mental health labels and matching case descriptions was accomplished with the assistance of a group of advanced-level interns in a program to train school psychologists ($n = 12$). Half of the trainees ($n = 6$) were randomly selected, given a group of mental health type labels taken from formal psychological nosologies or from common use, and asked to list behavioral manifestations or characteristics that would lead them to employ such a label. Behaviors and characteristics recurring most frequently were then incorporated into case descriptions by the investigator, and the other half of the group was asked to match the description to the correct label. For cross-validation purposes the same procedure was then repeated with the original group. The correct label was chosen unanimously in both groups for each of the four labels and matching case descriptions that were ultimately selected for use in the study:

1. Marginally retarded: This elementary school boy doesn't seem to have advanced as far as the others in the class in his school work. He usually takes somewhat longer than the others to finish his assignments. He seems to have particular difficulty with tasks like word problems in arithmetic, and his vocabulary is not as large as most of the other children's in the class.

2. Depressed neurotic: This junior high school boy is of average intelligence and usually completes his assignments satisfactorily. He is very quiet, however, and doesn't speak much in class or to his classmates. He frequently seems depressed and is often dissatisfied and discouraged about his own performance. Several of his teachers have occasionally noticed him crying. Even though his teachers seem satisfied with his work, he seems to always blame himself for not doing better. He seems to worry a great deal about his future, which he feels doesn't look very bright.

3. Emotionally disturbed: This elementary school boy frequently doesn't follow directions. He seems to have difficulty concentrating on his work. He occasionally will burst into laughter or tears or become very angry or upset for no apparent reason. He doesn't get along very well with the other children in the class.

4. Paranoid schizophrenic: This high school boy has great difficulty concentrating on his work. He is very suspicious of his teachers and the others in his classes. He stays to himself much of the time. When spoken to he doesn't seem to make sense at all. He states, for instance, that one of his teachers has put a mechanical device in his head that is controlling him. At times he has violent outbursts directed at one of his teachers or the other pupils.

Degree of incapacitation scale. In accordance with procedures presented by Edwards (1957), a modified Thurstone equal-appearing-interval scale was developed by the investigator to measure the subjects' assessments of the degree of psychological incapacitation manifested by the four hypothetical pupils. The 18 agree–disagree type statements that comprised the scale had been ranked by a group of expert judges, 11 counselors at the University of Maryland Counseling Center, along a 7-point scale representing a continuum from an attitude or view of a pupil as minimally incapacitated to an attitude of great incapacitation. As delineated by the investigator, a pupil would be considered incapacitated to the extent that (*a*) his behavior was regarded as beyond the normal range or "abnormal," (*b*) his prognosis was poor, (*c*) his condition was regarded as having a predetermined course and was embedded, chronic, and deep-rooted, (*d*) he was in need of specialist-type help, and (*e*) he had adverse effects upon his peers.

The item with the lowest score weight (1.42) was "It isn't very unusual for pupils of this age to act like this." The highest score weight (7.00) was assigned to the item, "A pupil like this should be institutionalized." An example of an item receiving an intermediate score weight (3.25) was "A pupil like this probably needs help through a difficult period." Interquartile ranges for the judges rankings for the 18 items ranged from .00 to 1.64.

A pilot study conducted with graduate students in a class to train remedial reading instructors ($n = 19$) yielded a reliability coefficient for the degree of incapacitation scale of .88 obtained by the test–retest method with a one-week interval.

Statistical Analysis

Data consisted of four scores for each subject on the degree of incapacitation scale—one for each of the four cases presented. Each of these scores was computed by taking the arithmetic mean of the scale values of the statements with which the subject agreed (Edwards, 1957).

Data were analyzed by means of an analysis of variance procedure employing a $4 \times 3 \times 2$ factorial design with repeated measures in one dimension, that of the four cases. Computerized analysis was employed (Dayton, 1971).

Results

Significant differences were found ($p < .001$) among mean scores on the degree of incapacitation scale for the three types of stimulus materials used—labels only, descriptions only, and descriptions with labels affixed ($F = 15.95$, $df = 2$). A post hoc analysis utilizing the Scheffé procedure indicated that the mean of the two conditions that employed labels combined was significantly higher than was the mean for the description-alone condition. No significant difference was found between the two conditions that utilized labels.

Although a significant F ratio was found for interaction effects between type of stimulus material and sex of pupil ($F = 3.91$, $df = 2$, $p < .05$), these differences were not located with the Scheffé test. This seeming discrepancy is probably attributable to the highly conservative nature of this procedure as contrasted to the analysis of variance test.

It was assumed at the outset that the four cases that were presented represented differing degrees of severity or incapacitation, and no hypotheses were formulated with regard to these differences. The fact that significant differences did, in fact, occur among these "known groups" ($F = 172.12$, conservative $df = 1$, $p < .001$) may, however, be cited as evidence in support of the validity of the measuring instrument. (Conservative degrees of freedom were employed because of failure to meet the homogeneity of covariance assumption.)

Significant interaction effects also occurred between type of stimulus material and the four cases ($F = 8.99$, $df = 2$, $p <$

.01). It was indicated in the post hoc analysis that the main treatment effects of type of stimulus material were accounted for by only two of the four cases, the marginally retarded and depressed neurotic. Main treatment effects were not significantly different for the emotionally disturbed and paranoid schizophrenic.

No significant overall difference was found between the means for the two pupil sexes, and there was no significant interaction effect between sex of pupil and the four cases.

Correlation coefficients between sex and length of teaching experience of the subjects and scores for the treatment variations were of low magnitude and not statistically significant.

Discussion

Effects of Labels

It should be observed that not only were the combined mean scores for the two conditions that utilized labels significantly higher than were those for the description-only condition, but that, when taken alone, the description-with-label-affixed mean was significantly higher than that of the description-alone condition. This finding is of particular interest when considered in the context of several theoretical positions that were reviewed. Ellis (1967) has suggested that the negative connotations of labels might be reduced if the labels were accompanied by operational definitions. If it can be assumed that the behavioral description that followed the label was, in effect, an operational definition, then Ellis' contention would not be supported by this study.

These findings, on the other hand, can be reconciled with several other theoretical positions. Scheff's (1966) proposition that it is the label that serves to stabilize what may be only a transitory condition might suggest that the label in this study added a connotation of stability, whereas without the label the condition might have been regarded as transitory or developmental in nature. Merton's (1957) contention that once a meaning has been ascribed to an event, it is interpreted in the light of this ascribed meaning would also seem applicable. The label in the description-with-label-affixed condition could assume its potency by ascribing meaning to the described behavior. It is finally suggested that the labeling effects might be interwoven, perhaps inextricably, with the stigmatizing effects of having sought help for mental health problems, a phenomenon that is supported in the literature (Phillips, 1963) and by recent political example. Attempts in subsequent research to unravel these effects might include setting up conditions in which the labeling is attributed to a mental health professional and to a non-professional source, such as a psychiatrist versus a teacher.

The significant interaction effects between the type of stimulus material and cases indicated that the labeling effects that were observed did not occur consistently across the four cases. Post hoc findings revealed significant differences in stimulus material effects for the depressed neurotic and marginally retarded cases, it is recalled, but not for the emotionally disturbed and paranoid schizophrenic cases. These results afford some possibility for conjecture about the qualities of labels that would render them more or less likely to produce the observed effects. One such quality might be the severity of the disorder described. The paranoid schizophrenic, for instance, is, of course, a more seriously disturbed individual than the depressed neurotic. Furthermore, an examination of the accompanying description of the paranoid schizophrenic reveals that even for this very serious disorder the symptoms described are exceedingly bizarre. The pupil is said, for instance, to make claims that the teacher has put a mechanical device in his head that is controlling him. In contrast, the descriptions of the marginally retarded and the depressed neurotic do not seem to present a very clear departure from normal. The marginally retarded child, for instance, is described as taking longer to finish his tasks. It might be hypothesized then that for less serious disorders in which the symptoms do not depart greatly from what is regarded as normal behavior, it is the label that serves to organize the somewhat ambiguous behavior into the abnormal category. It is the label

in other words, that ascribes an abnormal meaning to the symptoms. In instances in which the symptoms themselves show great deviance from what is regarded as normal or are perhaps the most severe of the symptoms associated with a given diagnostic category, the affixing of the label seems to have little effect. It remains, of course, for subsequent researchers to verify these observations with a larger and more systematically selected sampling of labels.

Sex Differences

No overall difference was observed between the mean scores for subjects given cases designated as boys and those given cases designated as girls. Significant interaction effects between type of stimulus material and sex of pupil, however, were observed. Further scrutiny revealed that no overall difference might be accounted for by the fact that for girls the scores under the behavioral-description-alone condition were lower and under the labeling conditions somewhat higher than they were for boys. These differences, then, would tend to cancel each other out in the overall analysis for sex.

Although the significant interaction effects between sex of pupil and type of stimulus material were not located by the Scheffé procedure, a trend was observed in the direction of greater labeling effects for girls. This trend was accounted for not so much because of large differences between boys and girls under the labeled conditions, but because under the behavioral-description-alone condition the girls were viewed as less incapacitated than were the boys. The trend in this direction would be consistent with the findings of Gurin, Veroff, and Feld (1960), who concluded that in our society there is greater tolerance of deviant behavior among females than among males. It might also be cited as consistent with the recent findings of Broverman, Broverman, Clarkson, Rosenkrantz, and Vogel (1971) that behavior that is regarded as "sick" for people in general is regarded as normal for females. This trend, if verified, would then seem of interest to those investigating the role of the schools in perpetuating sex role stereotypes.

It would not seem justifiable on the basis of one limited study to recommend that the use of current mental health nosologies be abandoned in the schools. Moreover, teachers and mental health practitioners may justify the claim that the communicative value of such labels overrides the biasing consideration. A recommendation that is seen as appropriate, however, is that the development of greater sensitivity and awareness of the potentially biasing effects of labels be incorporated into the preparation and in-service training of teachers and mental health professionals in the schools. An obvious alternative to the indiscriminate use of labels, and one that efforts already are being made to implement, is greater use of observations, descriptions, and recommendations for action at the behavioral level.

REFERENCES

Bentz, W., Edgerton, J. W., & Miller, Francis T. Perceptions of mental illness among public school teachers. *Sociology of Education,* 1969, **42,** 400–406.

Broverman, I. K., Broverman, D. M., Clarkson, F. E., Rosenkrantz, P. S., & Vogel, S. R. Sex-role stereotypes and clinical judgments of mental health. *Journal of Consulting and Clinical Psychology,* 1971, **70,** 1–7.

Cordaro, L., & Ison, J. Psychology of the scientist: X. Observer bias in the classical conditioning of the planarian. *Psychological Reports,* 1963, **13,** 787–789.

Dayton, C. W. Repeated measures program. (RMP-3) College Park: University of Maryland, Computer Science Center, 1971.

Edwards, A. E. *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts, 1957.

Ellis, A. Should some people be labeled mentally ill. *Journal of Consulting Psychology,* 1967, **31,** 435–446.

Gurin, G., Veroff, J., & Feld, S. *Americans view their mental health.* New York: Basic Books, 1960.

Ingraham, L. H., & Harrington, G. M. Psychology of the scientist: XVI. Experience of E as a variable in reducing experimenter bias. *Psychological Reports,* 1966, **19,** 455–461.

Merton, R. K. The self-fulfilling prophecy. In R. K. Merton (Ed.), *Social theory and social structure.* New York: Free Press, 1957.

Nunnally, J. E. *Popular conceptions of mental health.* New York: Holt, Rinehart & Winston, 1961.

Phillips, D. L. Rejection: A possible consequence of seeking help for mental disorders. *American Sociological Review,* 1963, **28,** 936–972.

ROSENTHAL, R., & JACOBSON, L. *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston, 1968.

SARBIN, T. R. On the futility of the proposition that some people be labeled mentally ill. *Journal of Consulting Psychology*, 1967, **31**, 447–453.

SCHEFF, T. J. *Being mentally ill: A sociological theory*. Chicago: Aldine, 1966.

STUKAT, K. *Suggestibility: A factorial and experimental analysis*. Stockholm, Sweden: Almqvist & Wiksell, 1958.

SZASZ, T. S. The myth of mental illness. *American Psychologist*, 1960, **15**, 113–118.

ULLMAN, L., & KRASNER, L. *A psychological approach to abnormal behavior*. Englewood Cliffs, N. J.: Prentice-Hall, 1969.

YAMAMOTO, K., & DIZNEY, H. Rejection of the mentally ill: A study of the attitudes of student teachers. *Journal of Counseling Psychology*, 1967, **31**, 216–237.

(Received February 7, 1973)

Manuscripts Accepted for Publication in the *Journal of Educational Psychology*

Learning from Prose Text: Effects of Readability Level, Inserted Question Difficulty, and Individual Differences. Jack H. Hiller (Southwest Regional Laboratory for Educational Research and Development, 4665 Lampson Avenue, Los Alamitos, California 90720).

Adjunct Questions and the Comprehension of Prose by Children. Ingrid Swenson and Raymond W. Kulhavy (Department of Educational Psychology, Arizona State University, Tempe, Arizona 85281).

Pseudo-Orthogonal and Other Analysis of Variance Designs Involving Individual Difference Variables. Lloyd G. Humprheys (Department of Psychology, University of Illinois, Champaign, Illinois 61820) and Allen Fleishman.

Morning-to-Afternoon Changes in Cognitive Performances and in the EEG. Edmund J. Mackenberg, Donald M. Broverman (Department of Psychology, The Commonwealth of Massachusetts Department of Mental Health, Worcester State Hospital, Worcester, Massachusetts 01613), William Vogel, and Edward L. Klaiber.

Effects of Feedback, Learner Control, and Cognitive Abilities on State Anxiety and Performance in a CAI task. Joe B. Hansen (Research and Evaluation, Education Service Center, Region XIII, 5604 Tracor Lane, Austin, Texas 78721).

A Follow-Up Study of Teacher Expectancy Effects. Edward J. O'Connell, Jerome B. Dusek (Psychology Department, 331 Huntington Hall, Syracuse University, Syracuse, New York 13210), and Richard J. Wheeler.

Variability in Children's Comprehension of Syntactic Structures. Alan M. Lesgold (Learning Research and Development Center, Mineral Industries Building, University of Pittsburgh, Pittsburgh, Pennsylvania 15260).

Task, Learner, and Treatment Variables in Instructional Design. John E. Rhetts, (Department of Educational Psychology, 330 Burton Hall, University of Minnesota, Minneapolis, Minnesota 55455).

Type and Frequency of Questions in Processing Textual Material. John P. Rickards (Educational Psychology and Research Section, SCC G #55, Purdue University, West Lafayette, Indiana 47907) and Francis J. Di Vesta.

Predictability of Students' Evaluations of College Teachers from Component Ratings. Grace French-Lazovik (Center for the Improvement of Teaching, University of Pittsburgh, Pittsburgh, Pennsylvania 19513).

Effects of Teacher Sex and Student Sex on the Evaluation of College Instructors. Patricia B. Elmore (Testing Center, Washington Square Building C, Southern Illinois University, Carbondale, Illinois 62901) and Karen A. LaPointe.

A Statistical Test of the Theoretical Model for the Representational Level of the ITPA. Eugene B. Doughtie (Department of Psychology, University of Houston, Cullen Boulevard, Houston, Texas 77004), James A. Wakefield, Jr., Raymond N. Sampson, and Herbert L. Alston.

Effects of Readability on Oral and Silent Reading Rates. Esther U. Coke (Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974).

Prediction of Productivity from Personality Test Scores. Charles F. Elton (Department of Higher and Adult Education, Room 111 Dickey Hall, University of Kentucky, Lexington, Kentucky 40506) and Harriet A. Rose.

Classificatory Behaviors of Low-Socioeconomic Status Children. Gordon K. Nelson (Department of Human Development, The Pennsylvania State University, Schuylkill Campus, State University—Schuylkill Haven, Pennsylvania 17972) and Herbert J. Klausmeier.

Unfamiliar Stimulus Terms in Children's Paired Associate Learning. Robert E. Davidson (Department of Educational Psychology, School of Education, University of Wisconsin—Madison, 1025 West Johnson Street, Madison, Wisconsin 52706), Sherry A. B. Perry, and Phoebe K. Baker.

Journal of Educational Psychology
1974, Vol. 66, No. 1, 123-128

PERSONALITY TRAITS ASSOCIATED WITH EFFECTIVE TEACHING IN RURAL AND URBAN SECONDARY SCHOOLS[1]

KENNETH D. MATTSSON[2]

University of Minnesota

The Cattell Sixteen Personality Factor Questionnaire was administered to 73 secondary students at the start of their student teaching. Near the end of the quarter their teaching effectiveness was assessed by administering the Hoyt-Grim Pupil Reaction Inventory to their pupils. Different patterns of personality traits associated with teaching effectiveness were found when subjects were grouped by level, by major field, and most distinctly, by size of community. Teaching effectiveness correlated in medium cities with Outgoing (A+), Emotionally Mature (C+), Trusting (L−), Confident (O−), Group Dependent (Q_2−), Relaxed (Q_4−), and Low Anxiety (I−) factors and in small towns with Sober (F−), Shy (H−), Sensitive (I+), Trusting (L−), and Introverted (II−) factors. On eight of the above factors there were direct and almost linear relationships between size of community and strength of the correlation with teaching effectiveness.

It would be pointless to dust off the historical record of man's interest in teachers' personality traits. But it would be useful to look at a few points of reference (and of curiosity) along the path over which we recently have come.

Seventy years ago a variety of what we now see as "quaint characterizations" were still being made about the teachers of children, characterizations such as the following written in 1902:

> Such constant, unrelieved association with juvenile life has its narrowing effect upon the mind of the master. After a time he loses his sense of perspective, attaches undue importance to trifles, and is disposed to judge of the world at large by the ethical and other standards of the schoolroom [Adams, p. 34].

Witty, a quarter of a century ago, observed that far too many teachers had developed an ideal of self that emphasized self-denial, abstinence, and deprivation. Not only could this result in the denial of many normal satisfactions and appetites, but also, "Such a personality tends to alienate children and young people [1947, p. 669]."

At the same time that Witty was making this observation in the name of the mental hygiene movement, Symonds (1947) was summarizing research in a pessimistic way, "there is no pattern of personality that will make the best teacher [p. 653]."

Not surprisingly, the difficulties experienced in the 1940s and 1950s did not extinguish research efforts. If anything, the study of teachers' personality traits has been revitalized by the development of new assessment instruments, and it now attracts the interest of an ever widening group of educational researchers. For example, Sandven (1969), at the University of Oslo, concluded from his study of 931 Norwegian teacher trainees, "It appears that social coreaction[3] represents a personality factor which in a constructive way has relevance to the role of teaching [p. 136]."

---

[1] This study is based on the author's doctoral dissertation submitted to the University of Minnesota.

Appreciation is extended to Gordon M. A. Mork for his guidance in the planning and execution of the study.

[2] Requests for reprints should be sent to Kenneth D. Mattsson, who is now at the School of Education, Mankato State College, Mankato, Minnesota 56001.

[3] Sandven uses the term social coreaction to mean the ability and tendency to react empathically and with sympathy.

Interest in teacher personality traits is still very much alive. The new concern for improved human relations within our schools is justification enough for greater effort to understand the personal qualities of those who are being called upon to effect these improvements.

The present study was an attempt to discover relationships between teacher personality traits as measured by the Sixteen Personality Factor Questionnaire (Sixteen PF), and success in classroom teaching as measured by the Hoyt–Grim Pupil Reaction Inventory (PRI). Since the available literature offered inconsistent and often contradictory findings, there was not sufficient basis for hypothesizing a specific set of personality traits associated with effective teaching. The hypothesis tested in this study was, therefore, a general one aimed at discovering, rather than verifying, patterns of traits and was stated as follows: There is no relationship between scores attained by student teachers on the Sixteen PF and scores attained on the PRI. This hypothesis was tested for student teachers grouped by grade level, major field, and size of community.

Method

Subjects

The subjects of this study were 73 Mankato State secondary school student teachers (49 men and 24 women) in the subject matter areas of language arts (23), mathematics (12), science (13), and social studies (26), who were assigned to full-time off-campus student teaching in 35 cooperating schools in southern Minnesota.

Instruments

The Sixteen PF, developed by Cattell, is an inventory of

> real, functionally unitary, and psychologically meaningful dimensions of personality, based upon factor analysis of raw data obtained from behavior ratings, objective personality tests and inventories, as well as from clinical and social reaction patterns as observed in a wide range of structured and unstructured situations [Cattell, 1957, p. 167].

Henjum (1967) correlated scores on the Sixteen PF with the PRI and found that successful junior high student teachers were extraverted, "well-adjusted," warm, friendly, and participating, while successful senior high student teachers were intelligent, enthusiastic, and practical. Other studies of teacher personality trai that used the Sixteen PF include Mitchell (196 Tarpey (1965), Morrison and Romoser (196 Warburton, Butcher, and Forrest (1963), Isaacso McKeachie, and Milholland (1963), Burdi (1963), and Kimmel (1964). An analysis of the studies, plus an additional five reviewed by He jum (1967), yields no consistent pattern of pe sonality traits associated with teaching effectiv ness. Part of the reason for a lack of consisten could be the failure of many researchers to repo their subjects' teaching fields, teaching levels, a sex. But, more importantly, none of the researc ers attempted to measure the very probable effe of the different cultural environments associat with rural communities on the one hand and large more urban communities on the other.

One very interesting study by Kerkman (196 did relate size of the community with pupil b haviors. He found that in large-city schools, p pils spent more time attending to the teacher a there was more organized group work and mo total teacher–pupil interaction, while classroom in small town schools were more heavily weighte on autonomous action of pupils, heterogeneity patterns, and the engaging of the teacher by t pupil.

In his review of research, Ryans (1960) co cluded that sex of teacher and size of communi were related to teacher effectiveness. The prese study was especially concerned with the latt factor.

The PRI is an inventory of pupil respons (agree, disagree, no opinion) to 200 statemen similar to those which pupils make about the teachers and classroom experiences (e.g., "T student teacher often doesn't seem to know h here." "We work in class, but we have fun too. The authors of the PRI found that it would yie scores independent of classifications by schoo grade-level, and subject matter field and that was "a source of unique information about teach behavior as manifested in its effect upon psych logical phenomena within pupils [Grim, Hoyt, Mayo, 1954, p. 84]."

The report by Henjum found the PRI to be satisfactory measure of classroom teaching effe tiveness and offers additional background info mation on the instrument (1967, p. 56).

Procedure

Forms A and B of the Sixteen PF were admini tered to the student teachers before they left th campus. The measure of classroom teaching effe tiveness, the PRI, was administered by the supe vising teachers to the pupils (about 3,500) of th student teachers during the last week of teachin Each subject was asked to name the best and th worst of his three classes and then the supervisin teacher was requested to administer the PRI t those two classes.

Product-moment correlation coefficients were computed between scores on the Sixteen PF and mean scores on the PRI for subjects grouped ac

cording to teaching level, subject matter area, and size of community.

Results

Scores on 20 personality factors (16 primary and 4 secondary factors of the Sixteen PF) were correlated with scores on the PRI for 73 junior and senior high school student teachers. In the following discussion of correlations a one- or two-word description of a Sixteen PF factor is followed by the factor letter, and since each factor is bipolar, a plus or minus sign is included.

As can be seen in Table 1, classroom effectiveness correlated significantly for junior high teachers with Conscientious (G+), Group-Dependent (Q_2-), and Relaxed (Q_4-) factors and for senior high teachers with the Group Dependent (Q_2-) factor.

When the subjects were grouped according to subject matter area, it was found that effectiveness in teaching language arts (English and speech) correlated significantly with the following factors of the Sixteen PF: Shy (H−), Trusting (L−), Relaxed (Q_4-), and Introverted (II−). One significant correlation each was found for math and science teachers, but with 20 correlations involved this could have happened by chance. Effectiveness in social studies correlated ($p < .01$) with the Group-Dependent (Q_2-) factor.

The findings for teaching level and subject matter area were strong enough to suggest recognizable patterns of personality traits that were associated with effective teaching in junior high and language arts but not in senior high nor in math, science, or social studies.

Much stronger patterns were found when the subjects were grouped according to the size of the community in which they taught. The size-of-community categories were (*a*)

TABLE 1

Correlations between Student Teacher Sixteen Personality Factor (pf) Scores and Mean Pupil Scores on the Hoyt-Grim Pupil Reaction Inventory by Teaching Level and Major Field

Sixteen PF factor description Low score (−) to high score (+)	Scale	Junior high (n = 36)	Senior high (n = 37)	Language arts (n = 22)	Math (n = 12)	Science (n = 13)	Social studies (n = 26)
Primary factor							
Reserved–Outgoing	A	.01	−.04	−.09	.42	−.19	−.08
Dull–Intelligent	B	−.10	.20	.05	−.20	.42	.13
Emotionally Immature–Mature	C	.08	−.09	.25	.40	.11	−.14
Submissive–Dominant	E	−.06	−.18	−.37	.21	−.28	−.13
Sober–Enthusiastic	F	−.17	−.07	−.26	.32	−.33	−.02
Expedient–Conscientious	G	.40*	−.04	.10	−.23	.04	.19
Shy–Venturesome	H	−.16	−.09	−.42*	.26	−.20	−.03
Tough Minded–Sensitive	I	.15	.13	.25	−.33	.60*	.14
Trusting–Suspicious	L	−.09	−.06	−.43*	−.08	.20	−.01
Conventional–Imaginative	M	.04	−.18	.29	−.13	−.19	−.14
Forthright–Shrewd	N	.00	−.03	−.03	−.02	.00	−.27
Confident–Apprehensive	O	−.15	−.20	−.21	−.66*	.17	−.18
Conservative–Experimenting	Q_1	.11	−.11	−.07	−.09	−.09	.08
Group Dependent–Self-sufficient	Q_2	−.38*	−.36*	−.09	−.34	−.45	−.49**
Casual–Controlled	Q_3	.20	.06	.33	−.00	.27	−.20
Relaxed–Tense	Q_4	−.37*	−.13	−.44*	−.32	−.17	−.04
Second-order factor							
Low Anxiety–High Anxiety	I	−.21	−.09	−.29	−.42	−.09	−.01
Introverted–Extraverted	II	−.13	−.10	−.48*	.41	−.29	−.01
Responsive–Tough Poise	III	−.14	−.10	−.24	.22	−.35	−.13
Dependent–Independent	IV	−.16	−.24	−.13	−.15	−.22	−.25

* $p < .05$.
** $p < .01$.

small town—population of fewer than 2,000; (*b*) *small city*—population of 2,000–7,499; (*c*) *medium city*—population of 7,500–30,000; and (*d*) *metropolitan*—(Minneapolis) and some suburban areas.

The correlations listed in Table 2 show distinct patterns of personality traits associated with effective teaching in medium cities and small towns but not in metropolitan or small-city schools.

In small towns, teaching effectiveness correlated significantly with Sober (F−), Shy (H−), Sensitive (I+), Trusting (L−), and Introverted (II−) factors. Teaching effectiveness in medium cities correlated significantly with the following factors: Outgoing (A+), Emotionally Mature (C+), Trusting (L−), Confident (O−), Group Dependent (Q_2−), Relaxed (Q_4−), and Low Anxiety (I−). On 8 of these 11 different factors the correlations were in opposite directions for the two groups of teachers.

As is made more apparent in Figure 1, the correlation coefficients on five of the factors (A, C, F, H, and Second-Order II) were strongly negative for the small town teachers and positive for the medium city teachers with the correlations for the small city teachers generally falling midway between those extremes. The direction of this relationship across the size-of-community dimension was just the opposite for Factors O, Q_4, and Second-Order I but even more pronounced in its nearly straight-line progression from the small town of fewer than 2,000 population to the medium city of 7,500–30,000.

Discussion

When the subjects of this study were grouped by teaching level and by subject matter area, the findings were minimally sufficient for the rejection of the null hypothesis. But when teachers in rural communities were compared with teachers in larger communities, the findings not only

TABLE 2

Correlations between Student-Teacher Sixteen Personality Factor (PF) Scores and Mean Pupil Scores on the Hoyt-Grim Pupil Reaction Inventory by Size of School Community

Sixteen PF factor description Low score (−) to high score (+)	Scale	Small town (n = 18)	Small city (n = 18)	Medium city (n = 18)	Metropolitan (n = 19)
Primary Factor					
Reserved–Outgoing	A	−.45	−.05	.53*	−.25
Dull–Intelligent	B	.18	−.06	.05	.25
Emotionally Immature–Mature	C	−.39	−.34	.67**	.03
Submissive–Dominant	E	−.19	.12	−.16	−.42
Sober–Enthusiastic	F	−.52*	.01	.19	−.10
Expedient–Conscientious	G	−.18	.24	.24	.38
Shy–Venturesome	H	−.62**	.14	.42	−.08
Tough Minded–Sensitive	I	.60**	.14	.21	−.11
Trusting–Suspicious	L	−.47*	.03	−.47*	.41
Conventional–Imaginative	M	−.03	−.07	−.03	.05
Forthright–Shrewd	N	−.22	−.27	.45	−.09
Confident–Apprehensive	O	.23	−.24	−.58*	−.17
Conservative–Experimenting	Q_1	−.04	−.15	−.05	.03
Group dependent–Self-sufficient	Q_2	−.05	−.50**	−.47*	−.36
Casual–Controlled	Q_3	−.08	−.10	.39	.32
Relaxed–Tense	Q_4	.03	−.32	−.70**	.01
Second-order factor					
Low Anxiety–High Anxiety	I	.30	−.17	−.75**	−.01
Introverted–Extraverted	II	−.63**	.15	.42	−.16
Responsive–Tough Poise	III	−.40	−.11	−.19	.05
Dependent–Independent	IV	.00	−.16	−.41	−.25

* $p < .05$.
** $p < .01$.

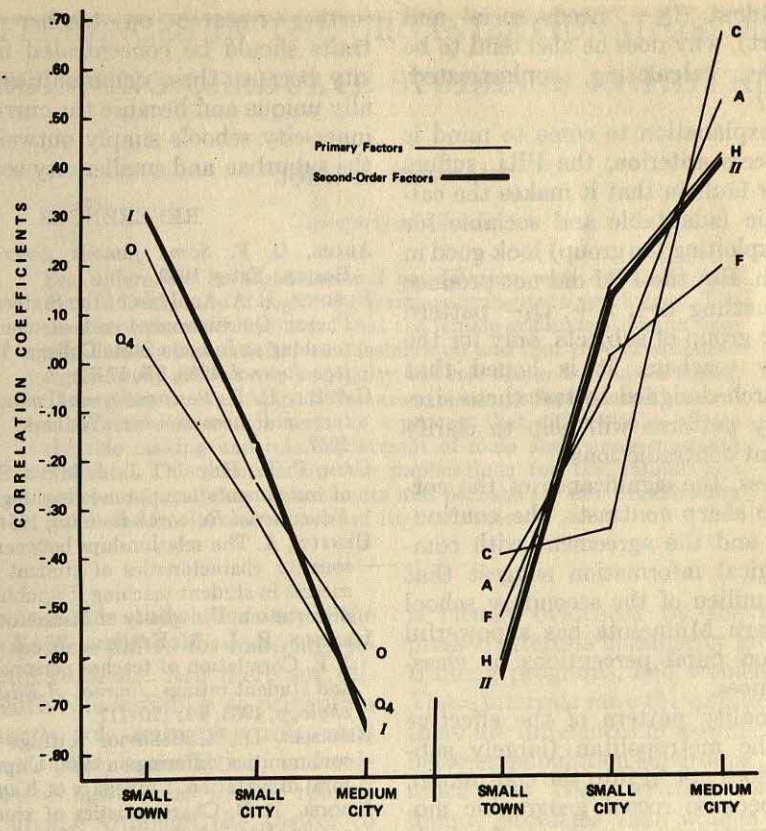


FIGURE 1. Relationships between selected correlation coefficients (Sixteen Personality Factor Questionnaire versus teaching effectiveness) and size of community.

necessitated the rejection of the null hypothesis but also gave substantial evidence that size of community was a more powerful factor than were teaching level or subject matter area in the relationship between successful teaching and personality traits of the teacher.

A careful analysis of the results reveals a distinct pattern of traits that appears to be related to successful teaching in the medium city and an equally strong but nearly opposite pattern related to successful teaching in the small town school whose student body is comprised largely of farm children.

For communities similar to those that hosted the student teachers used as the subjects of this study it can be inferred that the personality traits of successful beginning teachers in secondary academic areas range over a continuum from general introversion at one end to extraversion at the other. Starting with the effective student teacher in the small town who tends to be reserved, shy, sensitive, trusting, and introverted, the continuum encompasses the group-dependent small city teacher and terminates at the opposite end with the teacher in the medium city who tends to be more outgoing, mature, trusting, confident, relaxed, unanxious, and possibly extraverted. However, apparent contradictions in the small town and medium city patterns do appear in the findings. Why do these disparate types of teachers share the same value on Factor L, signifying that they are trusting, adaptable, free of jealousy, and easy to get along with? Any of the relationships could, of course, have occurred by chance, but did both groups appear to be trusting because the opposite value on L (indicating a tendency to be suspicious, self-opinionated, and hard to fool) is so foreign to effective teachers that trusting is the only possible result? And if the effective teacher in the medium city tends to be

group dependent (Q_2-, needs social and group support), why does he also tend to be shrewd (N+, calculating, sophisticated, hardheaded)?

The first explanation to come to mind is that the success criterion, the PRI, suffers from a major fault in that it makes the calculating cynic (adaptable and sociable for the sake of exploiting the group) look good in the classroom. But the PRI did not produce that same puzzling L−, N+, Q_2- pattern for any other group of subjects, only for the medium city teachers. It is hoped that further research designed to test these size-of-community patterns will help to clarify these apparent contradictions.

Nevertheless, the significance of the correlations, the sharp contrasts, the continuous pattern, and the agreement with common sociological information suggest that the cultural milieu of the secondary school in southwestern Minnesota has a powerful influence upon pupil perceptions of classroom experiences.

The personality pattern of the effective teacher in the metropolitan (largely suburban) area does not fit into the continuum, supposedly because recent geographic mobility has resulted in a mixture of people from large cities, small cities, and rural environments settling in the suburbs. The children of these people bring with them conflicting perceptions about teacher–pupil interactions in the classroom. As a corollary to this hypothesis, it can be expected that, in time, a new consensus of perceptions and expectations about teachers will synthesize in these areas of rapid change.

The findings of this study strongly suggest that certain cultural variables associated with size of community are strong determinants of the particular patterns of teachers' personality traits that are related to classroom effectiveness. The clear implication is that, similarly, there may be associated with inner-city neighborhoods certain cultural variables that operate, in a sense, to require particular patterns of personality traits for successful teaching or, at least, for the avoidance of failure.

In spite of the fact that research is generally more difficult to execute in the inner-city school than in the small town school, further research on teacher personality traits should be concentrated in the inner city because these communities are culturally unique and because the current needs of inner-city schools simply outweigh those of the suburban and smaller-city schools.

REFERENCES

Adams, O. F. *Some famous American schools.* Boston: Estes, 1902.

Burdick, L. A. Analysis of the Sixteen Personality Factor Questionnaire and elementary student teaching at Indiana State College. *Teacher's College Journal,* 1963, **35,** 57–59.

Cattell, R. B. *Personality and motivation structure and measurement.* Yonkers: World Book, 1957.

Grim, P. R., Hoyt, C. J., & Mayo, S. T. A study of instruments to appraise teaching competency. *Educational Research Bulletin,* 1954, **33,** 69–72.

Henjum, A. The relationships between certain personality characteristics of student teachers and success in student teaching. Unpublished doctoral dissertation, University of Minnesota, 1967.

Isaacson, R. L., McKeachie, W., & Milholland, J. E. Correlation of teacher personality variables and student ratings. *Journal of Educational Psychology,* 1963, **54,** 110–117.

Kerkman, D. H. Behavior settings at school in communities differing in size. Unpublished doctoral dissertation, University of Kansas, 1964.

Kimmel, P. B. Characteristics of student teachers as correlates of success in student teaching. Unpublished doctoral dissertation, Texas Technological College, 1964.

Mitchell, J. V. Personality characteristics associated with motives for entering teaching. *Phi Delta Kappan,* 1965, **46,** 529–532.

Morrison, W. L., & Romoser, R. C. Traditional classroom attitudes, the A.C.T. and the 16 P.F. *Journal of Educational Research,* 1967, **60,** 326–329.

Ryans, D. G. *Characteristics of teachers.* Washington, D. C.: American Council on Education, 1960.

Sandven, J. Do we select the right people for teacher training? *Pedagogica Europaea: The European Yearbook of Educational Research,* 1969, **5,** 113–136.

Symonds, P. Personality of the teacher. *Journal of Educational Research,* 1947, **40,** 652–661.

Tarpey, M. S. Personality factors in teacher trainee selection. *British Journal of Educational Psychology,* 1965, **35,** 140–149.

Warburton, F. W., Butcher, H. J., & Forrest, G. M. Predicting student performance in a university department of education. *British Journal of Educational Psychology,* 1963, **33,** 68–79.

Witty, P. An analysis of the personality traits of the effective teacher. *Journal of Educational Research,* 1947, **40,** 662–671.

(Received February 7, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 129-138

SEX DIFFERENCES IN THE ORGANIZATIONAL ASSIMILATION OF BEGINNING GRADUATE STUDENTS IN PSYCHOLOGY[1]

JOHN E. NEWMAN[2]

University of Illinois

Sex differences in organizational assimilation were examined in a longitudinal study of beginning graduate students in psychology. The results of the study indicated that the female graduate students were significantly less satisfied and assimilated and that they experienced significantly greater role ambiguity and role demands than the males. There were no significant differences in role performance. These results were discussed in terms of a process that differentially affects the role making and role adjustment of male and female graduate students. This information has implications for the assimilation policies of graduate departments and perhaps for any organization interested in the assimilation of its newcomers.

The purpose of this study was to see if the role-making process differs for male and female graduate students. Are there sex differences in such role variables as perception of role demands, role ambiguity, role satisfaction, role conflict, assimilation, and role performance for newcomers to the role of graduate student? Although the answer to this particular question is more than interesting, the importance of this type of research far exceeds the boundaries of the setting utilized in this study. A theory of role making and role adjustment is relevant to all individuals, in all roles, in all organizations. Therefore, the present research was undertaken primarily with the hope of contributing to the development of a general theory of role making and role adjustment which will serve as a basis for future research and for practical action.

The desirability of establishing a general theory of role making and role adjustment is clearly evidenced by the currently expressed interests in minority group selection, training programs, and women's liberation. These interests raise the question of whether there are differences in assimilation patterns between population subgroups. For example, much has been written concerning the unique problems that women face in academic settings (e.g., Astin, 1972; Kaley, 1971; Rossi, 1973), yet there is little empirical evidence to support any conclusive theory on why contemporary women find graduate education and professional occupations difficult to pursue. The present study focused on these problems from an organizational assimilation perspective and examined the assimilation of male and female newcomers in an educational organization.

What is organizational assimilation? The assimilation process was conceived as a role-making–role-adjustment process. What is the newcomer supposed to do? How does he find out? Who or what are the sources for role demands? What kind of and how much pressure is put on the new organizational member to do this or that role activity? How does he perceive and respond to these role demands? How is the new person's role defined? How much latitude does the newcomer have in defining his own role? What are the events and processes of role making and role adjustment?

The concept of *assimilation* derives its

[1] The author wishes to express his sincere appreciation to George B. Graen, Charles L. Hulin, Thomas W. Johnson, and J. B. Orris for their invaluable consultation and guidance and to the graduate students who generously participated in the study. The author would also like to thank Patrick R. Laughlin and Jeanne B. Herman for their comments on an earlier version of this paper.

[2] Requests for reprints should be sent to John E. Newman, who is now at 1 State Farm Plaza, State Farm Insurance Company, Bloomington, Illinois 61701.

meaning from the interrelationships among the set of role variables mentioned at the beginning of this article. The dynamic nature of the variables and their relationships make it unlikely that anyone ever becomes completely assimilated. (Note that complete assimilation may mean that one's role is totally and finally defined, accepted by all concerned, and performed as such with complete satisfaction and satisfactoriness.) However, it is assumed that one can measure the degree to which a newcomer or any organizational member feels assimilated. Note also that the conceptual framework and the methods involved in investigating the assimilation process offer an alternative approach to the study of organizational behavior, especially job satisfaction, job performance, and the elusive relationship between them.

Studies of the graduate student's role-making process, particularly as perceived by the graduate student, have been rare. Baird's (1969) article presented a good, brief review of the relevant research. Baird's own study of the role relations of graduate students is most related to the present research. Baird attempted an objective description of the social relations of graduate students. A questionnaire containing items describing the graduate student role was sent to 1,500 graduate students who returned 689 usable answer sheets. A factor analysis of these data extracted five factors to account for the role relations and general adaptation of students to graduate school: the extent of the student's involvement in graduate peer groups, the rigor of academic demands, the degree of ambiguity and conflict in professors' demands, the accessibility of the faculty, and the degree of tension the student experiences from these relations.

Baird's (1969) study and the present study have common conceptual origins in the theory of roles in complex organizations (Biddle & Thomas, 1966; Kahn, Wolfe, Quinn, Snoeck, & Rosenthal, 1964; Katz & Kahn, 1966). However, the present study goes beyond Baird's study in three important ways. First, this analysis provides a more dynamic picture. Baird provided a factorial description of the graduate student's role existing in but one time-slice of data. In an effort to examine the dynamic nature of role variables, the present longitudinal study included three time-slices of data and attempted not only to identify the factors or variables that enter the assimilation process, but also to monitor these variables and their interrelationships over time.

A second important addition concerns the nature of the focal subjects. Baird (1969) used a cross section of graduate students. By sampling graduate students at various degree stages, one is sampling students who have been in the organization for different lengths of time and, hence, one might expect differences in their extent of assimilation (or in their adaptation to the department as Baird might say). In essence, one might expect such variables relevant to role analysis as role ambiguity and role satisfaction to be affected by time spent in the organization. Therefore, Baird's results may be confounded by the differential amounts of time spent by his subjects in their respective graduate departments. In fact, Baird apparently realized this and, hence, claimed validity for some of his scales which had consistent relations with stage of graduate career. What effect time had on the other variables and scales is not known. Nor is it clear what effect the combining of data representing students at various career stages had on the generality of the five factors Baird extracted to account for the role relations of graduate students. One must also be careful if equating stage of career development with amount of time in the organization. For example, some students may have earned their master's degree in one year, other students in two or three years.

The third important difference is that the present study investigates sex differences, whereas Baird did not.

Therefore, the present study attempted to control and enhance the research design by (*a*) monitoring the role-related variables over time (time-sampling problem minimized), (*b*) focusing on an entire set (subject-sampling problem minimized) of first-year students beginning their graduate education (and their assimilation process) at the same point in time, and (*c*) ex-

amining sex differences in the assimilation process.

There was some basis for expecting sex differences in satisfaction. Smith, Kendall, and Hulin (1969) reported relatively consistent sex differences in job satisfaction as measured by the Job Description Index, with females being less satisfied. They suggested that such results may be related to possible differences in frames of reference and alternatives available for males and females. Hulin and Smith (1964) suggested that sex, per se, may not be the crucial factor which leads to high or low satisfaction. Instead, the entire set of variables which consistently covary with sex (e.g., pay, job level, promotion opportunities, and societal norms) may cause the differences in job satisfaction. It is important to note that the above findings were based on data obtained from men and women working in industrial settings.

In light of the previous research, the purpose of this study was to measure and monitor over time the role variables believed to be important in the assimilation process and to see if this role-making process differs for male and female graduate students. The hypotheses were as follows:

1. Females feel less satisfied with their role of graduate student than males.
2. Females perceive their role to be more ambiguous than males.
3. Females feel less assimilated than males.
4. Females perceive greater role demands than males.
5. Females experience greater role conflict than males.
6. Females have lower role performance evaluations (i.e., grade point averages) than males.

Method

Research Strategy and Subjects

It was believed that studying new graduate students during their first semester would be most appropriate for exploring the role-making and role adjustment process. The first few months of graduate school is a period when role demands are likely to be very salient and the new student's role-related attitudes and behaviors most subject to measurable change. Therefore, 44 first-semester graduate students in the psychology department of a large public university participated as the focal subjects (14 females, 30 males). There were no significant differences between females and males in terms of Graduate Record Examination scores (mathematics: $\bar{X}$ = 637 and 661, respectively; advanced test: $\bar{X}$ = 603 and 646, respectively) or in terms of quality of undergraduate school (Astin selectivity index: $\bar{X}$ = 61.4 for each group).

Instruments

A *taxonomy of role activities* was developed in the initial phase of the study by asking graduate students and faculty to what activities did first-semester graduate students allocate time and energy. Various samples of graduate students evaluated the resulting list of 100 role activities and a consensus was reached that 32 of the role activities were, indeed, valid activities of first-semester graduate students. Graduate students were then asked to assign to each role activity a *source* of pressure or demand for involvement in that activity. This resulted in three scales for the role activities: departmental, social (fellow students), and personal.

The *social* role activities scale consisted of the following items: (*a*) doing favors for my fellow students, (*b*) helping my fellow students with their problems, (*c*) listening to the complaints of my fellow students, (*d*) becoming an accepted member of the staff, (*e*) keeping good working relations with my fellow students, and (*f*) socializing with friends I have met through the department.

The *departmental* role activities scale contained these items: (*a*) learning basic methodology and how to approach problems, (*b*) learning what amount of work is required, (*c*) achieving high grades, (*d*) doing busy work, (*e*) working on papers and projects, (*f*) handling routine problems, and (*g*) learning the academic rules and regulations.

The *personal* role activities scale included the following: (*a*) planning my time to deal efficiently with peak work loads, (*b*) becoming psychologically adjusted to the graduate school routine, (*c*) finding out what courses are really like, (*d*) attending classes, (*e*) improving myself through my academic experiences, (*f*) getting things done in spite of the rules, (*g*) thinking about ways to improve my grades, (*h*) finding personal fulfillment as a graduate student, (*i*) getting myself going on my studies, (*j*) becoming master over my learning situation, (*k*) improving the quality of my work, (*l*) thinking about the bad aspects of getting an education, (*m*) making and keeping friendships, (*n*) learning who to go to for help on various problems, (*o*) doing my basic studying, (*p*) studying things in which I am really interested, (*q*) doing more than the minimum requirements, (*r*) reading texts and journals not assigned by the instructor, and (*s*) considering a change in discipline, speciality, or graduate school.

The *perceived role demands questionnaire* as-

sessed the magnitude of involvement demanded by each of the three sources of pressure (departmental, fellow students, personal). All role activities (i.e., all three scales) were listed for each source of pressure. The student responded to each role activity item by circling whether that source of pressure preferred that he spend more, the same, or less time than he presently spent on that activity. Items were scored 3 (more), 2 (same), and 1 (less). This questionnaire yielded nine scores—three scale scores for each of the three demand sources. A scale score was the sum of the scale's item scores. Because there were no sex differences in the perception of fellow students' demands, the present report concentrates its analysis on the perceived departmental and personal role demands and the departmental and personal role activities scales only.

Satisfaction indices. The work scale of the Job Description Index (Smith et al., 1969) was adapted (3 items were omitted) to assess the newcomer's satisfaction with his role. Items were scored following Smith et al. (1969, p. 79). The Hoppock (1935) index of overall job satisfaction was adapted to assess the newcomer's overall satisfaction with his role. This instrument contained four items, each with seven response alternatives. Each item was scored 1–7 (the higher score indicating greater satisfaction).

The *role ambiguity index* assessed how ambiguous the student perceived his role to be. This was a single 5-point scale administered at the end of the semester data collection. It was scored 1 (not ambiguous at all) to 5 (very ambiguous).

The *assimilation index* assessed how assimilated the student felt he was. This index contained one item with seven response alternatives and was scored 1 (I am not assimilated in the department at all) to 7 (I am completely assimilated in the department).

The above instruments were assembled into one questionnaire.

The *role performance indices*, first-semester and first-year grade point averages, were used to assess role performance. They had a possible range of from 1 (E) to 5 (A).

Conduct of the Study

The research design was one of repeated measures at three points during the students' first semester in graduate school. Data were collected at three 1-hour evening sessions—one during the second week of the semester, one at midsemester, and one during the last week of the semester.

Analysis

Univariate and multivariate analysis of variance techniques were used to assess the effect of sex, time, and their interaction on the role variables. Discriminant analysis was used to further understand the nature of the sex difference in role satisfaction. Correlational analysis was utilized to examine the interrelationships of the role variables.

Results

Role Satisfaction

Although satisfaction with the role declined over the semester for both groups, by the end of the first semester, females were significantly less satisfied with their role than were the males. Figure 1 illustrates this sex difference as measured by the adapted Hoppock (1935) index of overall role satisfaction. This finding lends support to Hypothesis 1.

To gain some understanding of the nature of this sex difference in satisfaction, the role descriptive items on the adapted Job Description Index were examined. Since there is a problem in interpreting differences between groups when each variable of a multivariate set is considered singly (see Tatsuoka, 1970), a multivariate discriminant analysis of the 15 Job Description Index role descriptive items was used to describe the group differences (Table 1). The estimated multivariate analogue of omega squared indicated that 59% of the variance in role satisfaction could be accounted for by group differences (i.e., sex differences). Interpretation of the discrim-

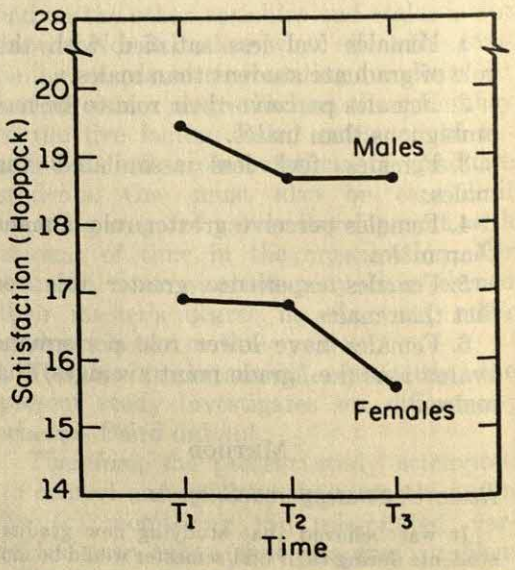


FIGURE 1. Sex differences in role satisfaction. (Significant sex difference on intrasession analysis of variance at T_3, $p < .05$. Significant sex difference on repeated-measures analysis of variance, $p < .05$. Abbreviations: T_1 = beginning of semester, T_2 = midsemester, and T_3 = end of semester.)

TABLE 1
RAW MEAN ENDORSEMENT VALUES AND DISCRIMINANT FUNCTION COEFFICIENTS FOR THE JOB DESCRIPTION INDEX ITEMS AT THE END OF THE SEMESTER

Job Description Index item	Raw $\bar{X}$		Normalized discriminant function coefficient	Scaled coefficient
	Female	Male		
Fascinating	.57	1.20	.31	2.35
Routine	.50	.80	−.09	−.68
Satisfying	.86	1.77	.29	2.48
Boring	1.93	1.87	.08	.74
Good	1.71	1.83	−.27	−2.33
Creative	.93	1.27	.22	1.94
Respected	.93	1.67	.17	1.32
Pleasant	1.50	1.53	.09	.76
Useful	2.29	1.97	−.22	−1.79
Tiresome	.93	.87	.03	.27
Healthful	.21	.73	.24	1.21
Challenging	1.64	1.63	−.48	−4.43
Frustrating	.57	.87	−.38	−3.11
Simple	2.21	2.40	.09	.65
Gives sense of accomplishment	.21	1.30	.40	2.67

Note. The discriminant function was significant beyond the .01 level.

inant function was based on the scaled variable loadings (Table 1). This interpretation suggests that those who score on the positive or high end of the role satisfaction dimension describe their role as fascinating, satisfying, and offering a sense of accomplishment, while those who score on the negative or low end of the dimension describe their role as frustrating. The group means on this dimension were .56 for the males and −.57 for the females. It should be noted that two other role descriptive items, "challenging" and "good," have large scaled coefficients but serve primarily as suppressor variables.

Perceived Role Ambiguity

As hypothesized, at the end of the semester, females perceived their role to be more ambiguous than did the males. The difference between the mean for the females ($\bar{X}$ = 3.71) and the mean for the males ($\bar{X}$ = 3.03) was significant at the .05 level. Role ambiguity was negatively correlated with role satisfaction (both indices) but only significantly so for the males (Table 2). (Note that all the correlations discussed in the "Results" refer to the end of the semester data collection and are presented in Table 2. Also, correlations for males and females were sometimes of the same magnitude but were only significant for the males because of the difference in the size of the two groups.)

Assimilation

At each data point, the graduate students were asked how assimilated in the department they felt they were. At semester's end, females were significantly ($p < .05$) less assimilated than the males who then felt about halfway assimilated (Figure 2). This important finding made Hypothesis 3 tenable and suggested that females were having a more difficult time in their role-making–role-adjustment process. Assimilation was positively correlated with overall role satisfaction for both males and females (Table 2). It is interesting to note, however, that assimilation correlated negatively with role ambiguity (−.57) for males but near zero (.08) for the females (Table 2).

Perceived Role Demands

This analysis was concerned with sex differences in the perception of demands for involvement in departmental and personal role activities. Of special interest were four distinct perceptions: (*a*) perceived departmental demand for involvement in departmental role activities (DD_d), (*b*) personal preference for involvement in departmental role activities (PP_d), (*c*) perceived departmental demand for involvement in personal role activities (DD_p), and (*d*) personal preference for involvement in personal role activities (PP_p).

Females perceived greater role demands (with respect to magnitude of involvement demanded) than did males throughout the first semester (Figures 3 and 4). Repeated-measures analysis of variance showed this sex difference to be very significant on two (DD_d and PP_p) of the four perceptions of role demands and relatively so on a third (PP_d; Table 3). These results offer partial support for Hypothesis 4.

Examination of the intrasession analyses of variance indicated that at the beginning of the semester there were no significant sex

TABLE 2

INTERCORRELATION MATRICES OF ROLE VARIABLES ASSESSED AT THE END OF THE SEMESTER FOR FEMALES AND MALES

Variable	Role											
	Demand				Conflict		Satisfaction		Ambiguity	Assimilation	Performance	
	1	2	3	4	5	6	7	8	9	10	11	12
Females[a]												
1. DD_d												
2. DD_p	.44											
3. PP_d	.18	.22										
4. PP_p	.05	.06	.47									
5. $\lvert RC_d \rvert$	.19	.14	−.48	−.20								
6. $\lvert RC_p \rvert$	.46	.53	.18	.33	.50							
7. Job Description Index Σ	.21	−.09	.05	.28	−.38	−.28						
8. Hoppock Σ	.11	.00	.07	.01	−.45	−.41	.82					
9. Assimilation	−.10	−.49	−.24	.03	.00	−.32	.47	.52				
10. Ambiguity	.27	−.24	−.09	.05	.37	.44	−.35	−.48	.08			
11. GPA_s	−.01	−.72	.10	−.12	−.47	−.59	.31	.21	.41	.26		
12. GPA_y	−.30	−.76	.11	.25	−.64	−.68	.44	.38	.48	−.09	.75	
Males[b]												
1. DD_d												
2. DD_p	.57											
3. PP_d	.24	.45										
4. PP_p	.54	.47	.12									
5. $\lvert RC_d \rvert$	.38	.17	−.52	.33								
6. $\lvert RC_p \rvert$	.08	−.39	−.06	.22	.15							
7. Job Description Index Σ	−.37	−.04	−.01	−.17	−.25	−.36						
8. Hoppock Σ	−.37	−.15	−.31	−.15	−.06	−.40	.80					
9. Assimilation	.01	−.17	−.27	−.18	.05	−.32	.31	.54				
10. Ambiguity	.06	.41	.29	.02	−.14	−.13	−.27	−.48	−.57			
11. GPA_s	−.24	−.40	−.11	−.14	−.21	−.23	.13	.18	.11	.01		
12. GPA_y	−.18	−.34	−.12	.02	−.14	−.18	.32	.32	.00	−.08	.76	

Note. Abbreviations: DD_d = perceived departmental demand for involvement in departmental role activities; DD_p = perceived departmental demand for involvement in personal role activities; PP_d = personal preference for involvement in departmental role activities; PP_p = personal preference for involvement in personal role activities; $\lvert RC_d \rvert$ = absolute role conflict with respect to departmental role activities; $\lvert RC_p \rvert$ = absolute role conflict with respect to personal role activities; GPA_s = first-semester grade point average; and GPA_y = first-year grade point average.

[a] $n = 14, r = .53, df = 12, p < .05$, two-tailed test.

[b] $n = 30, r = .36, df = 28, p < .05$, two-tailed test.

differences on the four role demand perceptions (Table 4). From examination of Figures 3 and 4 it appears that males and females did begin the semester with very similar overall perceptions (with the possible exception of PP_p) of departmental demands and personal preferences for involvement in the various role activities.

At midsemester the sex differences were in full bloom. Females perceived greater (now statistically significant) demands for involvement in role activities than did the males (Table 4). Male and female perceptions of role demands had diverged dramatically (Figures 3 and 4). Females had perceived an increase in the departmental demands for more involvement and had also increased their personal preference for involvement. The males' perceptions remained relatively stable (Figures 3 and 4).

Although the data (Figures 3 and 4) suggest that females continued to perceive greater role demands than the males at the end of the semester, the difference was statistically significant for only one (DD_d) of the four perceptions (Table 4).

In summary, the data indicate that the female graduate students did, indeed, experience greater role demands than did the males throughout the semester. Note also that both males and females perceived (*a*) greater *departmental* demand than personal preference for involvement in *departmental* role activities (Figure 3) and (*b*) greater *personal* preference than departmental demand for involvement in *personal* role activities (Figure 4). This represents an indication of the validity of the a priori role activities scales.

Perceived role demands (DD_d only) had a significant negative correlation with satisfaction for the males only (Table 2). The DD_p perception had a significant positive correlation with role ambiguity for the males (Table 2).

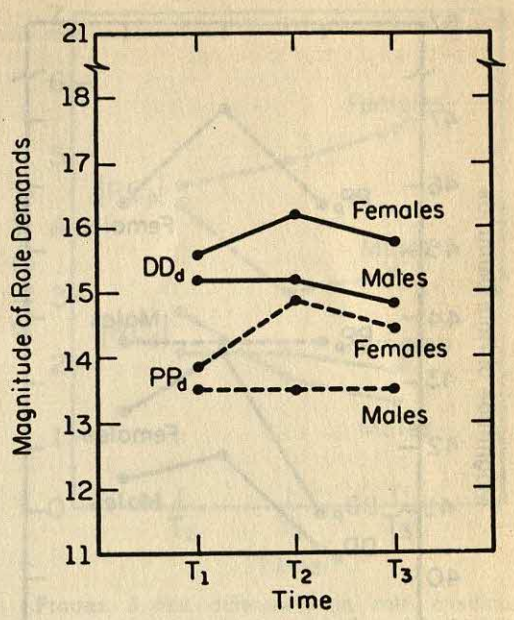


FIGURE 3. Sex differences in perception of role demands with respect to the departmental role activities scale. (Legend: —— = perceived departmental demand for involvement and - - - = personal preference (demand) for involvement. Abbreviations: T_1 = beginning of semester, T_2 = midsemester, and T_3 = end of semester.)

Role Conflict

Two measures of role conflict were derived by taking the absolute difference between the perceived departmental demands (DD) and personal demands (or preferences; PP) with respect to (*a*) the departmental role activities scale ($|DD_d - PP_d| = |RC_d|$) and (*b*) the personal role activities scale ($|DD_p - PP_p| = |RC_p|$). Note that this absolute measure of role conflict gives only the magnitude of conflict and not the direction. Although the data (Figure 5) indicate that females generally experienced a higher magnitude of role conflict than the males, the repeated-measures analysis of variance (Table 3) and intrasession analyses of variance (Table 4) revealed no statistically significant sex differences in role conflict for each of the above scales. Therefore, Hypothesis 5 failed to receive support.

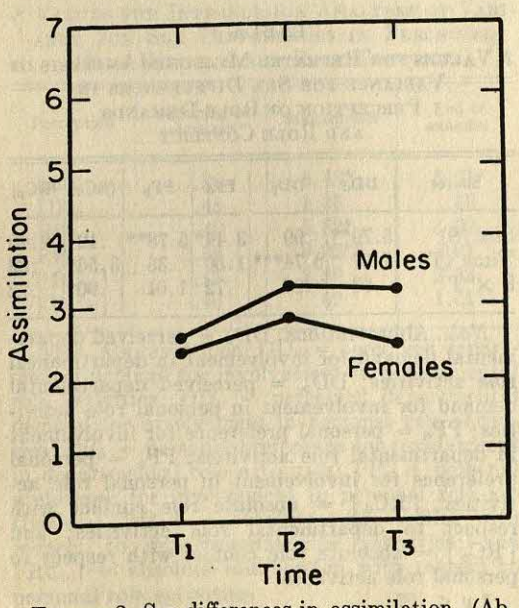


FIGURE 2. Sex differences in assimilation. (Abbreviations: T_1 = beginning of semester, T_2 = midsemester, and T_3 = end of semester.)

At the end of the semester, role conflict had a significant negative correlation with role satisfaction (both indices) for males (Table 2). Interestingly, both measures of role conflict had significant negative correlations with role performance (grade point averages) for the females (Table 2).

The personal role activities, and especially

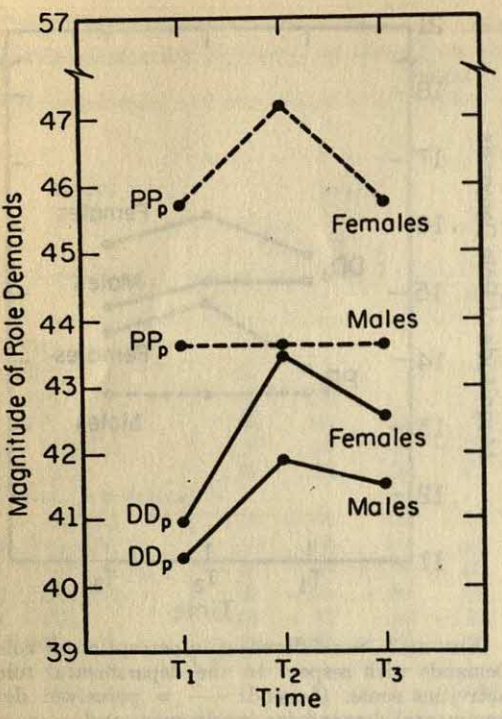


FIGURE 4. Sex differences in perception of role demands with respect to the personal role activities scale. (Legend: —— = perceived departmental demand for involvement and - - - = personal preference for involvement. Abbreviations: T_1 = beginning of semester, T_2 = mid-semester, and T_3 = end of semester.)

perceived departmental demand for involvement in them (DD_p), may have been a major source of the females' role adjustment problems. The DD_p significantly correlated .53 and −.39 with role conflict ($|RC_p|$) for females and males, respectively (Table 2). Seemingly, the greater the DD_p perceived by the females, the greater their role conflict, and the greater the DD_p perceived by the males, the less their role conflict. Remember also that DD_p increased significantly over the semester, especially for the females (see Table 4 and Figure 4).

Role Performance

Contrary to Hypothesis 6, there were no significant sex differences in role performance as indexed by grade point averages. The mean grade point averages for the first semester and for the first year, respectively, for the females were 4.73 and 4.80 and for the males were 4.85 and 4.84. Both $|RC_p|$ and DD_p were significantly negatively correlated with grade point average for females, while only DD_p was so related for the males (Table 2). (Note that correlations involving role performance may be more limited in this sample than in others because of the relatively restricted range of the grade point averages.)

Discussion and Summary

The broad purpose of this research was to investigate the assimilation process (i.e., role-making–role-adjustment process) of newcomers in an organization. This particular study examined considerable evidence for sex differences in the assimilation of students beginning graduate study in a psychology department. Females were significantly less satisfied and assimilated and they experienced significantly greater role ambiguity and role demands than the males. Although there was no significant sex difference in role performance (grade point average), the data overall suggest that the females may have had a more difficult time obtaining a level of performance comparable to the males.

These results suggest the operation of a process that differentially affects the role

TABLE 3

F VALUES FOR REPEATED-MEASURES ANALYSIS OF VARIANCE FOR SEX DIFFERENCES IN PERCEPTION OF ROLE DEMANDS AND ROLE CONFLICT

Source	DD_d	DD_p	PP_d	PP_p	$\|RC_d\|$	$\|RC_p\|$
Sex (S)	5.79**	.99	3.44*	5.78**	.12	2.40
Time (T)	1.27	6.74***	1.50	.38	3.56	.82
S × T	.93	.28	.72	1.01	.90	.72

Note. Abbreviations: DD_d = perceived departmental demand for involvement in departmental role activities; DD_p = perceived departmental demand for involvement in personal role activities; PP_d = personal preference for involvement in departmental role activities; PP_p = personal preference for involvement in personal role activities; $|RC_d|$ = absolute role conflict with respect to departmental role activities; and $|RC_p|$ = absolute role conflict with respect to personal role activities.

* $p < .07$.
** $p < .02$.
*** $p < .001$.

making and role adjustment of males and females during the first semester in a graduate program of psychology. The results do not specify in detail how the process operates, but they do give us some idea about what is happening and when it is happening. Seemingly, this process contributes to increasing the magnitude of the perceived role demands and role conflict for the females, while this is less so for the males. Perhaps, these perceptions then differentially affect their role satisfaction and role performance. The causal threads which are necessary to order these data and to explain these phenomena remain at a distance, but they are coming into view.

Although additional studies are needed to understand the assimilation process of graduate students in psychology departments, the differential reactions of females and males suggest that the current assimilating process may be somewhat more compatible with males than females. In light of the foregoing sex differences, perhaps graduate departments should reconsider the assimilation process and make necessary modifications.

In relation to Baird's (1969) study, the present study indicates the importance of complementing static, cross-sectional studies with longitudinal studies of homogeneous populations. The present study also makes it clear that future research on graduate student role relations must include an examination of potential sex differences.

In relation to current questions concerning the supposed dramatic change occurring in modern woman's role, the present study provided some evidence that, although some women have "come a long way, baby," getting there is not half the fun. That is, although females attained a level of role performance comparable to the males, they experienced less satisfaction and assimilation and more role demands, role ambiguity, and role conflict than did the males. This is essentially congruent with Kaley's (1971) conclusion that while there is some acceptance of the woman in a professional role on a theoretical level, there is rejection on an applied level.

Future research should not only further test the sex-difference hypotheses explored here, but should also consider research questions such as, Are there differences in

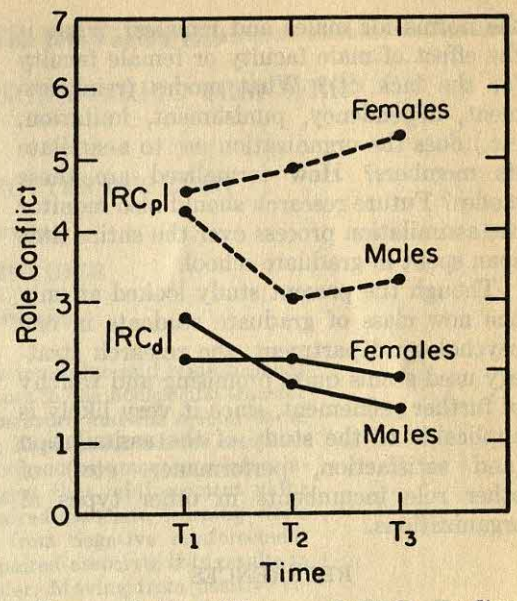


FIGURE 5. Sex differences in role conflict. (LEGEND: —— = absolute role conflict with respect to departmental role activities and - - - = absolute role conflict with respect to personal role activities. Abbreviations: T_1 = beginning of semester, T_2 = midsemester, and T_3 = end of semester.)

TABLE 4

F VALUES FOR INTRASESSION ANALYSES OF VARIANCE FOR SEX DIFFERENCES IN PERCEIVED ROLE DEMANDS AND ROLE CONFLICT

Perception	Beginning of semester	Midsemester	End of semester
DD_d	.35	5.48*	5.61*
DD_p	.48	1.35	.58
PP_d	.40	5.52*	3.09
PP_p	2.55	9.02**	2.91
$\lvert RC_p \rvert$	.17	2.89	2.33
$\lvert RC_d \rvert$	.40	.62	1.01

Note. Abbreviations: DD_d = perceived departmental demand for involvement in departmental role activities; DD_p = perceived departmental demand for involvement in personal role activities; PP_d = personal preference for involvement in departmental role activities; PP_p = personal preference for involvement in personal role activities; $\lvert RC_d \rvert$ = absolute role conflict with respect to departmental role activities; and $\lvert RC_p \rvert$ = absolute role conflict with respect to personal role activities.

* $p < .025$.

** $p < .005$.

the norms for males and females? What is the effect of male faculty or female faculty (or the lack of)? What modes (reinforcement, dependency, punishment, imitation, etc.) does the organization use to assimilate its members? How formalized are these modes? Future research should also monitor the assimilation process over the entire time span spent in graduate school.

Though the present study looked at only one new class of graduate students in one psychology department, the research strategy used seems quite promising and worthy of further refinement, since it very likely is applicable to the study of the assimilation (and satisfaction, performance, etc.) of other role incumbents in other types of organizations.

REFERENCES

Astin, H. S. Employment and career status of women psychologists. *American Psychologist*, 1972, **27**, 371–381.

Baird, L. L. A study of the role relations of graduate students. *Journal of Educational Psychology*, 1969, **60**, 15–21.

Biddle, B. J., & Thomas, E. J. (Eds.) *Role theory: Concepts and research*. New York: Wiley, 1966.

Hoppock, R. *Job satisfaction*. New York: Harper, 1935.

Hulin, C. L., & Smith, P. C. Sex differences in job satisfaction. *Journal of Applied Psychology*, 1964, **48**, 88–92.

Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoeck, J. D., & Rosenthal, R. A. *Organizational stress: Studies in role conflict and ambiguity*. New York: Wiley, 1964.

Kaley, M. M. Attitudes toward the dual role of the married professional woman. *American Psychologist*, 1971, **26**, 301–306.

Katz, D., & Kahn, R. L. *The social psychology of organizations*. New York: Wiley, 1966.

Newman, J. E. The assimilation of new graduate students in the department of psychology. Unpublished master's thesis, University of Illinois, 1971.

Rossi, A. S., & Calderwood, A. (Eds.) *Academic women on the move*. New York: Russell Sage Foundation, 1973.

Smith, P. C., Kendall, L. M., & Hulin, C. L. *The measurement of satisfaction in work and retirement*. Chicago: Rand McNally, 1969.

Tatsuoka, M. M. *Discriminant analysis*. Champaign, Ill.: Institute for Personality and Ability Testing, 1970.

Turner, R. H. Role-taking: Process versus conformity. In A. M. Rose (Ed.), *Human behavior and social processes*. Boston: Houghton-Mifflin, 1962.

(Received February 13, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 1, 139-151

FORMAL DISCIPLINE REVISITED:

AFFECTIVE ASSESSMENT AND NONSPECIFIC TRANSFER[1]

JOSEPH F. RYCHLAK[2] AND NGUYEN DUC TUAN

Purdue University

WILLIAM E. SCHNEIDER

Saint Louis University

Two meaningfulness measures, association value and reinforcement value, are considered as potential factors in the nonspecific transfer of learning; the first as a frequency measure and the second as an affective dimension of meaningfulness. Experiment 1 helps establish that reinforcement value cannot be reduced to nomothetic association value; Experiments 2 and 3 suggest that reinforcement value enters into nonspecific transfer in paired-associate learning while association value does not. Moving from negative reinforcement value to positive reinforcement value paired-associate lists results in the greatest positive nonspecific transfer. Moving from positive reinforcement value to negative reinforcement value paired-associate lists results in the least positive nonspecific transfer, and even in negative nonspecific transfer. Formal discipline theory may represent an inaccurate conceptualization of an accurate reinforcement value phenomenon.

The study of transfer in learning was originally stimulated by the subsequently discredited *formal discipline* thesis (Hall, 1971, pp. 357–358). Morgan (1906, p. 192), an advocate of this doctrine, held that the training of a student in Latin was invaluable in developing his "faculties" for comparison and generalization, which in turn would redound positively on his learning of English. As is well known, Thorndike and Woodworth (1901a, 1901b, 1901c) conducted a series of studies to set the formal discipline thesis into a decline from which it has never recovered. One of the main outcomes of the Thorndike-Woodworth argument was to solidify the construct of transfer in learning to a specificity thesis. Only those aspects of Latin which directly related to English as stimuli (as in word roots, etc.) could be given credit for improvements to be noted in a subject's performance across tasks. Unrelated tasks, such as learning arithmetic to improve grammar, could not be seen to enter into a transferred facilitation across areas of study (see Thorndike, 1914, p. 268).

This line of thought was carried into verbal learning as a concern with the specific characteristics of successive paired-associate lists. In order to evaluate the possible influence of factors *not* related to specific transfer effects, a control condition was employed of the order A–B, C–D. The specific factors of an A–B, A′–B sequence (similar stimuli and identical responses across two paired-associate lists) or an A–B, A–C sequence (identical stimuli and unrelated responses) could thus be compared with control conditions (A–B, C–D) in

[1] Experiment 2 and portions of Experiment 3 were supported by Grant 5-71-0039(509) from the U. S. Office of Education, Department of Health, Education, and Welfare awarded to Purdue University, with Joseph F. Rychlak as principal investigator. The opinions expressed herein, however, do not necessarily reflect the position or policy of the U. S. Office of Education, and no official endorsement by the U. S. Office of Education should be inferred. The authors would like to thank Marshall R. Schmidt and Sherry L. Felbain for their assistance in data collection.

[2] Requests for reprints should be sent to Joseph F. Rychlak, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

which no overlap in stimuli or responses took place.

Of course, this did not mean that improvement across the A–B, C–D lists was lacking. Subjects demonstrated what has since been termed a nonspecific (positive) transfer, which is usually defined as the improvement in performance to be noted across lists independent of list content (Kausler, 1966, p. 361). Note that though specific transfer (Thorndike, 1914) can be said to take a positive (improving) *or* negative (declining) direction, nonspecific transfer has been theoretically conceptualized as exclusively positive in nature. Theoretical explanations of specific transfer have followed mediation theory, with the constructs of stimulus discrimination and response integration given major emphasis (see, e.g., Kjeldergaard, 1968). Explanations of nonspecific transfer have usually emphasized the constructs of "warm-up" and "learning-to-learn" (see, e.g., Kintsch, 1970, pp. 35–36).

Aside from the belief in "faculties," which might be improved on the analogy of muscle training, the core distinction between formal discipline and the specificity thesis concerns just how much the subject brings to the task even when it is a novel one. Formal discipline held that the subject "comes at" the novel learning task with a more or less well-developed capacity to influence the outcome of what is eventually learned. If highly developed, this capacity leads to improvement; if poorly developed, a relative detriment in learning is likely to be the case (Roark, 1895, pp. 277–279). Specificity theory holds that only those antecedent experiences in identical or highly similar tasks can influence outcome in a seemingly novel task. If this previous experience is flatly contradictory to the present (novel) task, a decline in performance (negative transfer) will be the result. Certain "tricks of the trade" (learning to learn) incidental to the specific content of the material to be learned may also be learned; or, thanks to the motoric action of "warm-up" in the succession of tasks, a facilitation in rate may be induced. But nothing in the subject's conceptual, cognitive, or related intellectual capacities of a nonspecific character can be credited with inducing transfer effects.

The area of study in verbal learning which relates most directly to the question of what the subject might or might not bring to a novel situation is that of *meaning* (i.e., the point or significance which a sign or symbol carries for the subject) and its corollary metric of *meaningfulness* (i.e., the extent of the significance carried by the sign or symbol for the subject in question). The consonant-vowel-consonant trigram can be adapted as a controlled measure of meaningfulness. For example, even though one subject may think of POP as (meaning) a soft drink, another as (meaning) a parent, and a third as (meaning) an exploding noise, the fact that all three subjects do associate a meaning to this trigram makes it essentially meaningful for all concerned. Meaningfulness has been equated even though meaning has not. Material of this sort should be easier to learn than trigrams to which the subjects have no word associate.

A meaningfulness measure of this "frequency" type is termed (word) association value, and it can be measured *nomothetically* by finding the base rate of a trigram's word quality among a group of subjects (e.g., Archer, 1960). Trigrams with high base rates are thus considered more meaningful by the experimenter than those with low base rates. But association value can also be measured *idiographically* by having each subject rate whether or not he has a unique word associate to the trigrams he is to learn, regardless of the base rate for these materials. He may have an associate when most subjects do not or vice versa. In this case, meaningfulness is an "either–or" characteristic rather than a "high–low" characteristic of verbal materials.

Noble (1961) first raised the question of a nonspecific transfer in verbal learning, and he did so based on an earlier motor-learning study in which the specificity thesis had been given prominent consideration as a theoretical rationale. We refer here to the work of Jones and Bilodeau (1952), who had found that when the subject moved from a difficult motor task (cloverleaf pattern in a recessed tracking device) to a simple motor task (circle pattern), there was

greater positive transfer than vice versa. The question naturally arose, was this a specific transfer due to the fact that a circle practice effect was gotten via the subject's experience in drawing cloverleafs, or was this a form of nonspecific transfer? Noble felt it was probably the former, but in drawing an analogy between the difficulty levels of the motor tasks and the verbal learning of trigrams and quasi words, he wondered what would be the effect of moving from a difficult to an easy free-recall task, as compared to the opposite sequence. Would there be a greater (positive) nonspecific transfer across the former sequence?

If Latin is more difficult (less meaningful) for most students to learn than English (more meaningful), the implications of this research line for the formal discipline controversy are clear. Could it be that an observation made by the teacher when his students moved from such difficult areas of study to relatively easier areas resulted in the myth of formal discipline? Noble did not make this connection, of course. He had subjects memorize highly meaningful (easy) and less meaningful (hard) free-recall lists in contrasting orders. No demonstrable effect for his association value measure could be seen in the data for either sequence of lists (hard–easy or easy–hard). So confident was Noble (1961) in his research that he concluded, "Meaningfulness facilitates rate of acquisition but has no influence upon [nonspecific] transfer of training [p. 209]."

The subsequent research on this topic has been rather sparse. L'Abate (1962) contrasted subjects on high- versus low-nomothetic association value in a paired-associate task. He was primarily interested in the role of anxiety on learning, and his use of a rather involved measure of transfer makes his scanty findings suspect. He did show more nonspecific transfer for high association value than low association value across successive lists, but this was only for male subjects. Although a slight trend in favor of high association value leading to more nonspecific transfer over low association value was occasionally noted in a series of subsequent studies (many of which employed this tactic as a control condition; see the discussion above), in most cases the findings on association value meaningfulness in transfer are clearly negative (Dean & Kausler, 1964; Houston, 1965; Jung, 1963; Merikle & Battig, 1963; Stark, 1968). In no case was a "negative" nonspecific transfer found.

Findings from a more recent study (Rychlak & Tobin, 1971) were to shift the question of nonspecific transfer in verbal learning to an alternate dimension of meaningfulness. We refer here to *reinforcement value*, which is a rating of likability done on learnable materials preliminary to entering them into tasks for the subject to acquire (Rychlak, 1966). The reinforcement value dimension of meaningfulness is viewed as an affective assessment, with the construct "affect" viewed as the *psychological* evaluation made of the experience by the subject. Often this assessment is made on the basis of positive or negative physiological reactions known as emotions, but it is not necessary to be feeling an emotion (physical) in order to make an affective (psychological) assessment of liking or disliking.

A continuing effort has been made to prove that association value and reinforcement value are independent dimensions of meaningfulness. For example, even though Pearsonian correlations of from .54 to .84 can be found between nomothetic ratings of association value and reinforcement value, when these same data are reanalyzed idiographically via chi-square, it is found that a serious confounding of association value and reinforcement value ratings takes place only at the lower levels of nomothetic ratings for trigrams (i.e., where few subjects have a word associate to the trigram in question and the ratings for reinforcement value are highly likely to be negative; see Tenbrunsel, Nishball, & Rychlak, 1968). At the 80% to 100% nomothetic association value base-rate levels, where trigrams are easy to identify as having literal word meanings, only about one trigram in five is rated identically by the subject as to association value (yes, it is a word *versus* no, it is not a word) and reinforcement value (liked *versus* disliked, respectively). One-week test–retest reliability for rein-

forcement value is comparable to association value (in the .90s nomothetically, with about 35 out of 100 subjects changing their ratings when considered idiographically). Experimental evidence of the independence of reinforcement value from association value in learning was readily obtained. The more rapid acquisition of liked over disliked materials (termed the positive reinforcement value effect) was demonstrated without statistical interaction with association value across all levels of nomothetic association value and, within any such level, was statistically independent of idiographic association value measures as well (Abramson, Tasto, & Rychlak, 1969).

The positive reinforcement value facilitation has been found in paired-associate learning and in free-recall studies, using either mixed or unmixed lists. Such learning effects have also been found in the recognition of designs, abstract paintings, and pictures of human faces which have been prerated for reinforcement value (Rychlak, Galster, & McFarland, 1972). The age range of subjects in which reinforcement value effects on learning have been noted spans first grade through past 50 years (Rychlak & Saluri, in press). An interesting finding on abnormals (schizophrenics) is that such subjects either collapse the positive reinforcement value effect to insignificance or reverse it entirely (i.e., actually learn their disliked materials more rapidly than their liked; Rychlak, McKee, Schneider, & Abramson, 1971). Finally, black females have been shown to learn more along reinforcement value than association value relative to white females, who show the opposite predilection (Rychlak, Hewitt, & Hewitt, 1973).

In the Rychlak and Tobin (1971) study, overachieving male high school students (relatively low IQ, high grades) were contrasted with a matched group of underachievers (relatively high IQ, low grades). Since this was an unmixed paired-associate-lists study, it was necessary to counterbalance for the order in which subjects took their tasks (liked–disliked versus disliked–liked). This was tantamount to an A–B, C–D "control" condition in the typical transfer study discussed above. The general findings were that the underachiever reflected a larger positive reinforcement value effect than the overachiever. But in counterbalancing lists, it was inadvertently discovered that subjects who learned disliked lists before liked lists improved significantly more on their second list than did subjects who moved from a liked to a disliked list. Indeed, the lower IQ subjects of the sample actually reflected a decline in performance—that is, a negative (nonspecific) transfer—across the liked–disliked sequence.

If one now equates negative reinforcement value materials with "difficult to learn" and positive reinforcement value materials with "easy to learn," the findings of Jones and Bilodeau (1952) which Noble (1961) was unable to extend to verbal materials via the association value dimension *are* given support. Could an affective assessment of meaningfulness be the historical source of formal discipline theory? A test of the hypothesis that reinforcement value contributes to nonspecific transfer while association value does not was called for. As an additional development of this research line, a recent criticism made of reinforcement value procedures was also put to test. Three studies are presented: the first takes up the criticism referred to and the last two confront association value and reinforcement value in nonspecific transfer paired-associate tasks.

Experiment 1

Method

The following hypotheses were tested:

1. Ratings made by the subject of the sort "yes, it [the trigram] is word related and I like it" increase in proportion as successively higher levels of nomothetic association value are sampled.
2. Ratings of "yes, it's word related but I dislike it" *also* increase in proportion as successively higher levels of nomothetic association value are sampled.
3. Ratings of "no, it's not word related and I dislike it" increase in proportion as successively lower levels of nomothetic association value are sampled.
4. Ratings of "no, it's not word related but I like it" *also* increase in proportion as successively lower levels of nomothetic association value are sampled.

Rationale. Previous work on reinforcement value had established that the rating of "like versus dislike" employed to operationalize affective assessments could *not* be reduced to either a nomothetic or an idiographic (frequency of) association value measure (Rychlak et al., 1971). Subjects did not associate more words to liked than to disliked trigrams, no matter what nomothetic level these trigrams were taken from, so that a "frequency" explanation of reinforcement value (as due to the number of association-value-related associations) could not be defended. The word associates proffered by the subjects did fall properly into line, however, in that disliked trigrams suggested disliked word associates and liked trigrams suggested liked word associates. A criticism which has been raised since this earlier finding was reported suggests that just because the subject rates "liked" to a nomothetically low-ranging trigram (in association value) for which he has no word associate, it does not follow that all nomothetic association value factors are removed from potential influence on his subsequent learning effort. The trigram might thus appear to be one of "no, it's not a word but I like it," when in fact the actual word quality for this trigram is much higher in a "true" base-rate sense than this idiographic rating would suggest.

In other words, this criticism holds that reinforcement value measures capitalize on the error of measurement, selecting out those trigrams which have a covertly higher association value than the base-rate norms (gleaned from other subjects' judgments) currently reflect for positive reinforcement value and those which have a lower association value than these norms now reflect for negative reinforcement value. Learning rate differentials across positive and negative reinforcement value would thus be explained on the basis of this covert artifact.

To counter this reasonable charge, it is necessary to show that positive reinforcement value ratings of trigrams which are judged by subjects as idiographically without association value will be *more frequent* at the lower levels of nomothetic (base-rate) association value norms than they will at the higher levels (Hypothesis 4). Conversely, trigrams that are judged as negative in reinforcement value but as having idiographic word quality should be *more frequent* at the higher levels of nomothetic (base-rate) association value norms than they are at the lower levels (Hypothesis 2). In other words, the idiographic reinforcement value judgment will be shown to go against the drift of nomothetic association value levels even as idiographic association value ratings are performing true to expectation. The association value ratings will fall or rise idiographically in line with the levels of nomothetic association value from which all trigrams are selected, but the reinforcement value ratings will show a reverse tendency. How is it possible to conceive of reinforcement value as a covert nomothetic association value measure—capitalizing on the error of measurement—if the crucial negative reinforcement value ratings are more frequent at a higher nomothetic association value level and the crucial positive reinforcement value ratings are more frequent at a lower nomothetic association value level?

The remaining two combinations of idiographic reinforcement value and association value are expected to show a common proportion change across the levels of nomothetic association value. That is, subjects can be expected to find it difficult to like what they have no word meaning for, so this rating should increase in proportion as the nomothetic levels of association value drop off (Hypothesis 3). Conversely, when a trigram has word quality, it is more probable that the subject will find a liked meaning in his associations, so this rating should increase in proportion as the nomothetic levels of association value increase (Hypothesis 1). Hence, the crucial hypotheses which help establish that reinforcement value is not a covert measure of either idiographic or nomothetic association value are Hypotheses 2 and 4.

Subjects. Two hundred and fifty-six undergraduate college students (120 males, 136 females) were divided by random assignment into five groups for the rating procedure. There were approximately 50 subjects per group, divided equally by sex.

Procedure. Five levels of nomothetic association value were delineated, based upon Archer's (1960) base-rate norms for consonant-vowel-consonant trigrams: 0%–20%, 21%–40%, 41%–60%, 61%–80%, and 81%–100%. Two hundred trigrams were randomly selected from each of these base-rate levels (discarding those which had reflected a sex difference in Archer's survey) and then were administered to one of the five groups of subjects. Each subject thus rated 200 trigrams. In the usual rating procedure (see Experiments 2 and 3), association value and reinforcement value are assessed separately. However, in the present study a subject was asked to make *both* association value and reinforcement value ratings at the same time.

Each subject was asked to make two concurrent decisions: (*a*) to judge whether or not the trigram "looks like a word, sounds like a word, or can it be used in a sentence?" (after Archer, 1960) and (*b*) to judge whether he liked or disliked the trigram on the basis of how it "sounds" as he read it "aloud" to himself. Thus, a subject checked one of the following four alternatives for each of the 200 trigrams he was exposed to via special rating forms:

___ yes–like (yes, word related, and liked the sound)
___ yes–dislike
___ no–like
___ no–dislike

This rating procedure was repeated after a 48-hour period had elapsed.[3] Only trigrams which

[3] This procedure has since been shown to result

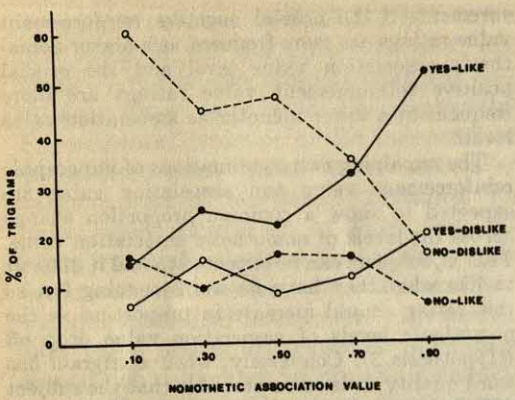


FIGURE 1. The mean percentage of trigram ratings in four scoring categories (yes–like, yes–dislike, no–dislike, and no–like) as reported by 50 subjects at each of five nomothetic association value levels are presented.

had been rated identically on both occasions were used in the data array and statistical analysis.

Results

A *score* for each of the four rating categories was determined for each subject. This reflected the percentage of the total number of reliable trigrams he had assigned to the rating category in question. Figure 1 presents the mean percentage of trigrams in a scoring category, distributed across the five levels of nomothetic association value.

Figure 1 will be reviewed by experimental hypothesis. Note that the yes–like (Hypothesis 1) distribution rises with increasing levels of nomothetic association value (as measured across the five experimental groups of roughly 50 subjects per group). The yes–dislike proportion (Hypothesis 2) also shows a rising trend with increasing nomothetic association value. The no–dislike (Hypothesis 3) rating is considerably larger in proportion score at the lower levels of nomothetic association value than it is at the higher levels. Finally, the no–like (Hypothesis 4) rating drops rather decidedly at the highest level of nomothetic association value in comparison to the level it had achieved at the lowest level.

When these proportion scores were expressed as linear functions of the five mid-range values of nomothetic association value (graphed as .10, .30, .50, .70, and .90), the following regression equations were found: $y_{yl} = .39x + .12$ for yes–like; $y_{yd} = .11x + .09$ for yes–dislike; $y_{nd} = -.45x + .61$ for no–dislike; and $y_{nl} = -.05x + .17$ for no–like. The correlations associated with these regression lines ($N = 250$) are .46 ($p < .001$), .22 ($p < .001$), $-.47$ ($p < .001$), and $-.09$ ($p < .07$), respectively. Thus, the slopes of these graphed lines were significantly different from zero in all but the no–like condition (which approximated significance and on a one-tailed presumption could be said to have reached significance). It would appear from Figure 1 that any criticism suggesting that reinforcement value is an artifact of nomothetic association value would be difficult to sustain.

EXPERIMENT 2

Method

The following hypotheses were tested:

1. Idiographically assessed association value and reinforcement value have significant but independent (main) effects on acquisition in paired-associate tasks.
2. Idiographically assessed association value fails to have interlist effects and reinforcement value succeeds in having interlist effects in a nonspecific transfer design having the paired-associate characteristics of A–B, C–D.
3. When the subject moves from a disliked to a liked task, the greatest positive nonspecific transfer occurs.
4. When the subject moves from a liked to a disliked task, the least positive transfer occurs.

Rationale. The first two hypotheses embody Noble's (1961) dictum, and there is ample evidence from earlier research on idiographic association value and reinforcement value that Hypothesis 1 is to be expected (Abramson et al., 1969; Rychlak et al., 1973). Hypotheses 2, 3, and 4 flow from the previous work on association value in transfer and the findings of Rychlak and Tobin (1971) discussed in the introduction.

Subjects and procedure. Subjects were 64 (32 males, 32 females) volunteer college students, primarily from the freshman and sophomore class levels. They made association value and reinforcement value ratings in a group and were then tested individually in the paired-associate task. A subject was administered a mimeographed rating form on which 140 trigrams from the middle ranges of Archer's (1960) nomothetic association values (40%–70%) were printed (see Rychlak, 1966, for the initial work done on these trigrams).

in the usual reinforcement value learning effects. Indeed, it has become the standard rating procedure since the present studies were conducted.

On the first two occasions, the subject was administered the trigrams with 48 hours intervening and asked to answer yes or no to the question, "does the trigram look like a word, sound like a word, remind you of a word, or can you use it in a sentence?" (after Archer, 1960). Those trigrams rated *yes* on both occasions were taken as meaningful to the subject in the sense of association value, and those rated *no* were considered less meaningful or unmeaningful.

One week following the association value ratings, the subject was again administered the 140 trigrams and asked to rate them for reinforcement value, based upon how they "looked" and "sounded" to him as he read them to himself. In this case, the more usual procedure was followed which involves a 4-point rating of like much, like slightly, dislike slightly, and dislike much (see Experiment 1 and Footnote 3). The convention followed in scoring for reinforcement value is to administer trigrams on two occasions with 48 hours intervening. Only those trigrams rated liked or disliked (preferably, "much") on both scoring occasions are then employed to construct the subject's paired-associate lists. The unreliable association value ratings were also ignored by the experimenters. A series of eight conditions were then arrayed by the experimenters, which perfectly counterbalanced association value and reinforcement value across two entirely different six-pair paired-associate lists. A mixed design of three between-subjects variables and one within-subjects variable was the result (2^4 factorial). The variables were Sex × Reinforcement Value × Association Value × Lists.

Subjects were randomly assigned to these eight conditions. In List 1, a subject might thus learn six pairs having the idiographic quality of association-value–no, reinforcement-value–liked and then follow this with a list of association-value–no, reinforcement-value–disliked. In this case, positive reinforcement value would have been removed from the materials to be learned in the second list. In another instance, the subject might move across lists having the characteristics of association-value–no, reinforcement-value–disliked to association-value–yes, reinforcement-value–disliked. In this case, association value meaningfulness would have been "added into" the materials to be learned. It should be emphasized that *both* paired-associate members had identical meaningfulness qualities; that is, so-called stimulus versus response meaningfulness was not under study.

The lists were presented by memory drum, with three different orders of lists to counter serial learning cues and a 4-second exposure time. Method of anticipation was followed, with two correct anticipations for the entire list taken as learning criterion. Total testing time occupied from 20 minutes to 1 hour. In both Experiments 2 and 3, the following steps were taken to counter the effects of letter overlap, rhyming, and so forth on rate of acquisition: phonetically similar trigrams (TUP and CUP) were not used, and if the two trigrams to be paired began with the same consonant or the first trigram ended with the consonant of the second, these pairings were also discarded (e.g., CUL and CAL or TUC and CUP).

Results

A 2^4 factorial analysis of variance was computed with three between factors (sex, association value, reinforcement value) and one within factor (lists). Although females tended to reach criterion faster than males, there was no significant sex difference. A significant main effect was found for association value, with means (and standard deviations) as follows: association-value–yes, 12.11 (6.83); association-value–no, 14.72 (7.19) ($F = 6.44$, $df = 1/56$, $p < .01$). A reinforcement value main effect was also apparent, as follows: reinforcement-value–liked, 11.77 (5.65); reinforcement-value–disliked, 15.06 (7.57) ($F = 10.26$, $df = 1/56$, $p = .002$). A pronounced nonspecific (positive) transfer effect was evident, as follows: List 1, 16.31 (7.51); List 2, 10.52 (4.62) ($F = 29.64$, $df = 1/56$, $p < .001$). There was no interaction between lists and either association value or reinforcement value, but this was due to the nature of this analysis of variance which would require a differential effect between lists to reflect a significant finding. In short, the first analysis of variance demonstrated that association value and reinforcement value each had effects in *both* lists and that these effects did not differ between lists. Since there was no interaction between association value and reinforcement value, the independence of the two measures of meaningfulness was completely supported. Hypothesis 1 was thus entirely substantiated.

In order to test the remaining hypotheses, two analyses of variance were run which collapsed one measure of meaningfulness on the other, while the latter was being assessed sequentially, as follows: ++, +−, −+, −−. This called for a 2 × 4 × 2 (Sex × Sequences × Lists) analysis of variance with the first two factors between conditions and the last a within-subjects condition. For example, the experimenters could first assess the role of association-value–yes (+) to association-value–yes (+), association-value–yes (+) to association-value–no (−), association-value–no (−) to

TABLE 1

MEANS AND STANDARD DEVIATIONS OF ASSOCIATION VALUE AND REINFORCEMENT VALUE MEANINGFULNESS SCORES ACROSS LISTS

Order of meaningfulness across Lists 1 and 2	Meaningfulness							
	Association value				Reinforcement value[a]			
	List 1		List 2		List 1		List 2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
+ to +	15.37	8.01	8.69	3.57	14.63	5.12	8.44	3.74
+ to −	15.06	5.33	11.88	4.53	14.31	6.19	11.19	5.28
− to +	16.69	6.71	9.31	3.69	19.18	10.00	9.68	4.37
− to −	18.13	10.47	12.19	5.41	17.13	6.74	12.75	3.75

[a] Order and list interaction, $p < .03$.

association-value–yes (+), and association-value–no (−) to association-value–no (−) with reinforcement value held constant across these sequences. Then a second run of the 2 × 4 × 2 analysis of variance would establish these identical patterns of liked versus disliked reinforcement value with association value held constant.

Table 1 contains the means and standard deviations of association value and reinforcement value across lists in terms of the order sequences outlined. The crucial test of Hypothesis 2 depended upon interaction effects between the four order sequences and the two lists. This interaction failed to reach significance for association value ($F = 1.69$, $df = 3/56$). However, it *did* achieve significance on reinforcement value ($F = 3.15$, $df = 3/56$, $p = .03$). There were no sex main effects or interactions. Note that the condition accounting for maximum (positive) transfer was when the subject moved from negative reinforcement value to positive reinforcement value and that the condition reflecting the least nonspecific transfer was when the subject moved from positive reinforcement value to negative reinforcement value. This is a cross-validation of Rychlak and Tobin (1971). Hypotheses 2, 3, and 4 are thus supported in the data analyses.

EXPERIMENT 3

Method

The following hypotheses were tested:

1. Reinforcement value shows nonspecific transfer effects across *four* successive paired-associate lists.

2. Moving from positive reinforcement value to negative reinforcement value results in the least (positive) transfer regardless of association value effects.

3. Moving from negative reinforcement value to positive reinforcement value results in the greatest (positive) nonspecific transfer regardless of association value influences in the paired-associate lists.

Rationale. If reinforcement value has interlist effects across two lists, it seems plausible to expect such effects across several lists. The order of these interlist effects should follow Experiment 2 as regards maximum and minimum gains in paired-associate acquisition. The relatively unimportant role for idiographic association value found in Experiment 2 was expected to continue.

Subjects. The subjects were 16 adults (8 males, 8 females), ranging in educational background from the 8th to the 12th grade. They were screened with the Shipley (1946) intelligence scale to rule out subnormal intelligence (IQ below 90). Job backgrounds for these subjects included factory work, domestic, janitorial, and practical nurse activities. All subjects fell in the "upper-lower" socioeconomic class category, based on the Hollingshead and Redlich (1958) index. These subjects were recompensed for their participation in Experiment 3.

Procedure. The strategy of Experiment 3 was to arrange a series of four paired-associate lists which varied sequentially in idiographic association value and reinforcement value in such a way as to generate interlist effects according to the hypotheses under test. Perfect counterbalancing of all possible meaningfulness combinations was not attained, nor was it considered necessary for a testing of the hypotheses. The first step involved having the subject idiographically rate 200 trigrams selected from roughly the 70% level of Archer's (1960) norms for both association value

and reinforcement value (exactly as in Experiment 2). Pretesting had established this to be the most satisfactory level of difficulty (nomothetic association value) for subjects of this general background.

Next, four different six-pair lists were arrayed for the subject, having the characteristics of (A) association-value–yes, positive reinforcement value (liked); (B) association-value–no, positive reinforcement value; (C) association-value–yes, negative reinforcement value (disliked); and (D) association-value–no, negative reinforcement value. Both paired-associate members of each of these lists had the same association-value–reinforcement-value meaningfulness combinations; that is, so-called stimulus versus response meaningfulness was not at issue. Hence there were four unmixed lists, administered to the subject in one of *four* sequences (see Figure 2). The males and females were assigned to the conditions randomly within the sex breakdown, so that two members of each sex were performing in each of the four sequences. Lists were presented via memory drum, with three randomized orders used to counteract serial learning cues and a 4-second exposure time. Subjects were given three "warm-up" trials on a practice list. Method of anticipation was followed, and learning criterion was again two successive correct anticipations on the entire six-pair list. Experimental periods took from 45 minutes to 1½ hours, with an ample "break" given between lists for the subject to go to the lavatory, smoke a cigarette, and so forth.

Results

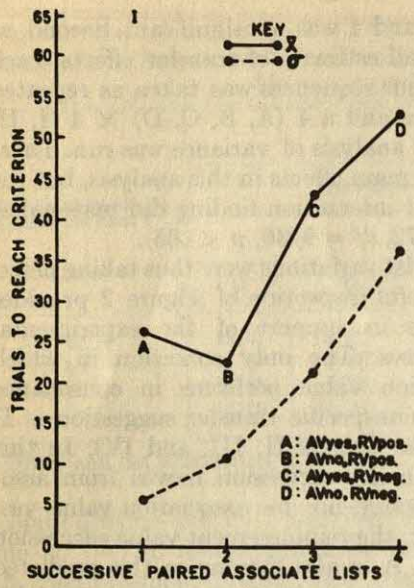
Figure 2 presents the means and standard deviations of trials to reach criterion in graph form, broken down into the four association-value–reinforcement-value sequence conditions (I, II, III, IV). The four graphs of Figure 2 thus trace the *interlist* performance of four subjects (two males, two females) across varying sequences of association-value–reinforcement-value meaningfulness (symbolized by the above A, B, C, D coding). Nonspecific transfer was assessed statistically in two ways.

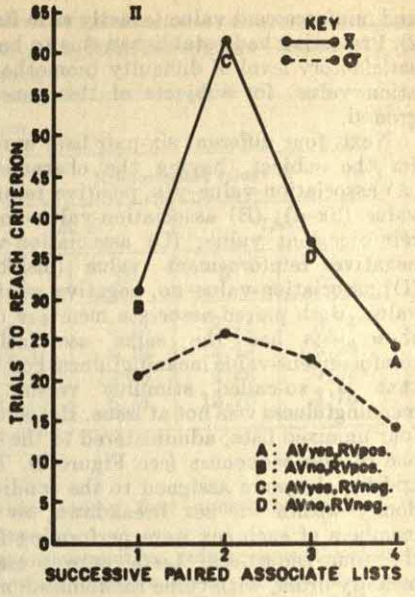
First, difference scores were obtained for each subject based on his successive performances across his four lists. These data were then submitted to a 2 (Sex) × 4 (A, B, C, D) factorial analysis of variance, with the following results: (*a*) there was no main sex or interaction effect; (*b*) there was a difference score main effect between Lists 1 and 2 ($F = 4.69$, $df = 3/8$, $p < .05$); (*c*) there was a difference score main effect between Lists 2 and 3 ($F = 6.32$, $df = 3/8$, $p < .05$); and (*d*) the difference score main effect between Lists 3 and 4 was nonsignificant. Second, as an overall estimate of transfer effects, each of the four sequences was taken as repeated measures and a 4 (A, B, C, D) × 4 (I, II, III, IV) analysis of variance was run. There were no main effects in this analysis, but the expected interaction finding did materialize ($F = 2.72$, $df = 9/36$, $p < .05$).

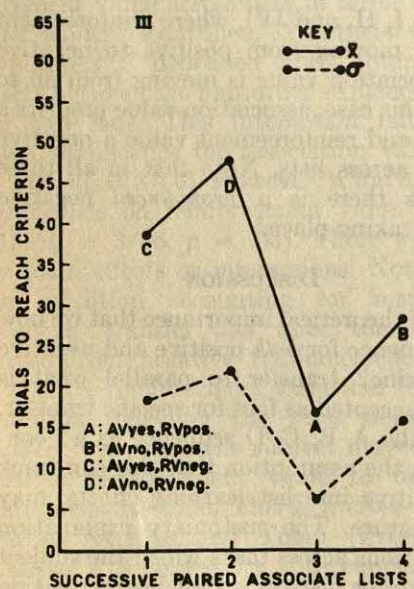
Interlist variations were thus taking place, and careful inspection of Figure 2 provides evidence in support of the experimental hypotheses. The only condition in which association value performs in consonance with a nonspecific transfer suggestion is D to A (see Graphs II, III, and IV). In this case, the list succession moves from association-value–no to association-value–yes; however, the reinforcement value succession of D to A is *also* moving in the predicted negative to positive (maximum positive transfer) direction. The A to B (Graphs I, III, and IV) and C to D (Graphs I, II, and III) sequences on association value predict negative transfer but are not consistently found to be so. The major confrontation of association value and reinforcement value comes within the B to C sequence of lists (Graphs I, II, and IV), where reinforcement value is moving from positive to negative and association value is moving from no to yes. In this case, association value predicts a positive and reinforcement value a negative transfer across lists. Note that in all three instances there is a *pronounced* negative transfer taking place.

Discussion

It is of theoretical importance that we now have evidence for *both* positive and negative "nonspecific" transfer to parallel what is already accepted as fact for specific transfer. To use the A–B, C–D sequence as a "control," on the assumption that this can track *only* positive interlist learning effects, may be premature. The customary explanation for variations across tasks which the subject either likes or dislikes is that his motivation level fluctuates accordingly. The subjects of Experiment 3 might thus have simply been "trying harder" on those lists which they liked, and this accounted for the differential rate of learning across the four lists. However there was no reason (or "reward") for these







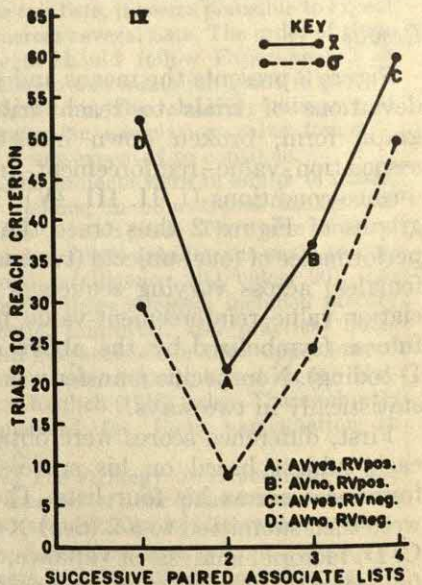


FIGURE 2. The means (solid line) and standard deviations (broken line) of 16 subjects are presented in separate graphs of 4 subjects each according to a different sequence of four paired-associate lists. (The conditions I, II, III, and IV demonstrate the overall order of reaching criterion, and the association-value–reinforcement-value combinations to be found in a specific list along the abscissa are symbolized by A, B, C, and D. The standard deviations are scaled on the ordinate raw score units for ease in presentation. Abbreviations: AV = association value; RV = reinforcement value; pos. = positive; and neg. = negative.)

subjects to delay completion of the total experiment. Payment was made at the close of study. Interviews with the subjects following the task convinced the experimenters that none were aware of the experimental manipulation. Though all trigrams were from the 70% nomothetic level of association value, the subjects definitely felt that some lists were more difficult than others. Experiment 1 counters any suggestion that the experimenters might have somehow preselected aberrant nomothetic trigrams for these lists. To argue that the subjects are motivated to try harder—and hence do better—on easier tasks results in a circularity.

Experiment 2 is clearly the stronger of the two studies on transfer, due to the perfect counterbalancing of conditions achieved. Here again, in light of the findings of Experiment 1, it does not seem profitable to begin speculating on covert nomothetic association value factors which somehow crept in to determine the results of this confrontation between idiographic association value and reinforcement value. It is disconcerting to traditional learning theory that a two-way transfer takes place across lists which have nothing observably "in common" (no stimulus or response similarities). Yet, this seems to be the case in the realm of affective assessment and—as suggested in the introduction—it is what unites reinforcement value and formal discipline theory. Both conceptions view the performing subject as having an *active* contribution to make to the learning process.

This side of formal discipline theory is often overlooked. Due to the stereotype of a mentalistic "muscle training," too many psychologists assume that the advocates of formal discipline believed that effort expended in a task to develop "faculties" could be passively induced. Actually, the proponents of formal discipline held that a teacher could *do nothing* for the student, who had to bring about his own (faculty) development (see, e.g., Roark, 1895, p. 278). The analogy to muscles was drawn along "active, intentional use" lines and *not* along "power through practice " lines. A student could be shown the way by the teacher and hopefully be aided when questions arose, but simply to force meaningless repetitions on him in hopes of furthering his "mental muscles" was not a tenet of formal discipline theory. In this context, it is possible to see how a reinforcement value effect across the less liked but dutifully performed Latin lesson might have seemed to generate by comparison a more talented performance in the better liked English or history lesson to follow. This reinforcement value sequencing of lesson content could actually be tested in a classroom situation.

It is a challenge to psychology to find out precisely what constitutes judgments of "hard" versus "easy" material. Informal observation of students at all school levels suggests to the experimenters that an affective factor is surely prominent in such statements—one that is more than simply a motivational variation of the type discussed above. This performance differential is often tied to conceptions such as "working at one's pace" or "self-selection of study topics" in the open classroom and so forth. It could be that the ultimate benefit of these so-called liberal classroom tactics is that they afford the student an opportunity to delve lightly into a disliked study topic (possibly by way of a teacher's suggestion or example) and then to devote considerable subsequent effort to a preferred study task (thus maximizing positive nonspecific transfer for the latter, which continues to be "easy" as greater gains are made). Or, the student may devote his entire time to various liked subjects (a sequence of learning which also results in pronounced positive transfer). As the total school effort is arranged to maximize facilitation of learning, the child may even come to reevaluate his formerly disliked subjects and place them in a more favorable light.

To undertake an extended discussion of the theory which sustains reinforcement value research would go beyond the aims of the present paper.[4] Suffice to say, it is assumed that in the act of assessing items which are learnable (trigrams, words, pic-

[4] Requests for a complete theoretical statement of the "logical learning theory" underwriting reinforcement value research should be sent to the first author.

tures, etc.), the subject equates the affective assessment he makes of himself with the affective assessment he makes of these materials—in a *tautological* fashion (liked item to liked self, or disliked item to disliked self). Such meaningful relations are thought to function "within time," so that explanations relying on practice, rehearsal, learning how to learn, or warm-up are not considered viable accounts of the reinforcement value phenomenon. These concepts are best left to explain association value learning effects. Reinforcement value indicates what will be found meaningful—hence "known"—by the subject as he "comes at" a novel experience. Knowing an item in this sense means it will reach learning criterion in a practice series before the unknown items do. But this does not make the seeming acquisition over time a function of repetition. In one sense, and to a certain extent, the "acquisition" is made possible due to the preliminary grasp (knowing) which the subject brings to the situation beforehand as an affective assessment.

REFERENCES

ABRAMSON, Y., TASTO, D. L., & RYCHLAK, J. F. Nomothetic vs. idiographic influences of association value and reinforcement value on learning. *Journal of Experimental Research in Personality*, 1969, **4**, 65–71.

ARCHER, E. J. Re-evaluation of the meaningfulness of all possible CVC trigrams. *Psychological Monographs*, 1960, **74** (10, Whole No. 497).

DEAN, M. G., & KAUSLER, D. H. Degree of first-list learning and stimulus meaningfulness as related to transfer in the A-B, C-B paradigm. *Journal of Verbal Learning and Verbal Behavior*, 1964, **3**, 330–334.

HALL, J. F. *Verbal learning and retention.* New York: Lippincott, 1971.

HOLLINGSHEAD, A. B., & REDLICH, F. C. *Social class and mental illness.* New York: Wiley, 1958.

HOUSTON, J. P. Supplementary report: Transfer as a function of meaningfulness. *Journal of Verbal Learning and Verbal Behavior*, 1965, **4**, 339–340.

JONES, E. I., & BILODEAU, E. A. *Differential transfer of training between motor tasks of different difficulty.* (USAF Research Bulletin 52-35) Lackland Air Force Base, Tex.: Human Resources Center, 1952.

JUNG, J. Effects of response meaningfulness (*m*) on transfer of training under two different paradigms. *Journal of Experimental Psychology*, 1963, **65**, 377–384.

KAUSLER, D. H. *Readings in verbal learning.* New York: Wiley, 1966.

KINTSCH, W. *Learning, memory, and conceptual processes.* New York: Wiley, 1970.

KJELDERGAARD, P. M. Transfer and mediation in verbal learning. In T. R. Dixon & D. L. Horton (Eds.), *Verbal behavior and general behavior theory.* Englewood Cliffs, N. J.: Prentice-Hall, 1968.

L'ABATE, L. Transfer of learning with differences in associative value and in manifest anxiety. *American Journal of Psychology*, 1962, **75**, 251–285.

MERIKLE, P. M., & BATTIG, W. F. Transfer of training as a function of experimental paradigm and meaningfulness. *Journal of Verbal Learning and Verbal Behavior*, 1963, **2**, 485–488.

MORGAN, C. L. *Psychology for teachers.* New York: Scribner, 1906.

NOBLE, C. E. Meaningfulness (*m*) and transfer phenomena in serial verbal learning. *The Journal of Psychology*, 1961, **52**, 201–210.

ROARK, R. N. *Psychology in education.* New York: American Book, 1895.

RYCHLAK, J. F. Reinforcement value: A suggested idiographic, intensity dimension of meaningfulness for the personality theorist. *Journal of Personality*, 1966, **34**, 311–335.

RYCHLAK, J. F., GALSTER, J., & MCFARLAND, K. K. The role of affective assessment in associative learning: From designs and CVC trigrams to faces and names. *Journal of Experimental Research in Personality*, 1972, **6**, 186–194.

RYCHLAK, J. F., HEWITT, C. W., & HEWITT, J. Affective evaluation, word quality, and the verbal learning styles of black versus white junior college females. *Journal of Personality and Social Psychology*, 1973, **27**, 248–255.

RYCHLAK, J. F., MCKEE, D. B., SCHNEIDER, W. E., & ABRAMSON, Y. Affective evaluation in the verbal learning styles of normals and abnormals. *Journal of Abnormal Psychology*, 1971, **77**, 247–257.

RYCHLAK, J. F., & SALURI, R. E. Affective assessment in the learning of names by fifth- and sixth-grade children. *Journal of Genetic Psychology*, in press.

RYCHLAK, J. F., & TOBIN, T. J. Order effects in the affective learning styles of overachievers and underachievers. *Journal of Educational Psychology*, 1971, **62**, 141–147.

SHIPLEY, W. C. *Shipley-Institute of Living scale for measuring intellectual impairment.* Hartford, Conn.: Institute of Living, 1946. (Manual)

STARK, K. Transfer from serial to paired-associate learning: A reappraisal. *Journal of Verbal Learning and Verbal Behavior*, 1968, **7**, 20–30.

TENBRUNSEL, T. W., NISHBALL, E. R., & RYCHLAK, J. F. The idiographic relationship between association value and reinforcement value, and the nature of meaning. *Journal of Personality*, 1968, **36**, 126–137.

THORNDIKE, E. L. *Educational psychology: Briefer course.* New York: Columbia University Press, 1914.

THORNDIKE, E. L., & WOODWORTH, R. S. The in-

fluence of improvement in one mental function upon the efficiency of other functions. I. *Psychological Review,* 1901, **8,** 247–261. (a)

THORNDIKE, E. L., & WOODWORTH, R. S. The influence of improvement in one mental function upon the efficiency of other functions. II. The stimulation of magnitudes. *Psychological Review,* 1901, **8,** 384–395. (b)

THORNDIKE, E. L., & WOODWORTH, R. S. The influence of improvement in one mental function upon the efficiency of other functions. III. Functions involving attention, observation, and discrimination. *Psychological Review,* 1901, **8,** 553–564. (c)

(Received February 19, 1973)

ANNOUNCEMENT

In addition to the members of the Advisory Board, the *Journal of Educational Psychology* asks other psychologists to aid in the review of manuscripts. In this way we have been able to achieve a broader review base. The Editors wish to thank the following people for their generous assistance to the Journal.

SCARVIA B. ANDERSON, *Educational Testing Service*
JOAN BARATZ, *Educational Study Center*
DANIEL E. BERLYNE, *University of Toronto*
BENJAMIN S. BLOOM, *University of Chicago*
JERE E. BROPHY, *University of Texas, Austin*
JOHN B. CARROLL, *Educational Testing Service*
JOHN E. CENTRA, *Educational Testing Service*
ROBERT S. DAVIDSON, *University of Wisconsin—Madison*
JOEL DAVITZ, *Teachers College, Columbia University*
DAVID H. DODD, *University of Utah*
JAN D. DUKER, *Teachers College, Columbia University*
RUTH B. EKSTROM, *Educational Testing Service*
* JANET ELASHOFF, *Stanford University*
GERARD FANELLI, *Newark State College*
LEONARD S. FELDT, *University of Iowa*
PAUL A. GAMES, *Pennsylvania State University*
LILA GLEITMAN, *University of Pennsylvania*
JOHN T. GUTHRIE, *Johns Hopkins University*
ROBERT HAYGOOD, *Arizona State University*
ALBERT N. HIERONYMOUS, *University of Iowa*
JAMES G. HOLLAND, *University of Pittsburgh*
HARRY LEVIN, *Massachusetts Institute of Technology*
* JOEL R. LEVIN, *University of Wisconsin—Madison*
DONALD MEDLEY, *University of Virginia*
ROBERT J. MENGES, *Northwestern University*
LEO A. MUNDAY, *American College Testing Program*
CALVIN F. NODINE, *Temple University*
DANIEL OGILVIE, *Livingston College, Rutgers University*
JOAN D. RITTENHOUSE, *National Research Council*
* ROBERT R. ROSENTHAL, *Harvard University*
DONALD B. RUBIN, *Educational Testing Service*
MARSHALL SANBORN, *University of Wisconsin—Madison*
THOMAS J. SHUELL, *State University of New York at Buffalo*
RICHARD E. SNOW, *Stanford University*
MARVIN SONTAG, *Teachers College, Columbia University*
JUDITH V. TORNEY, *University of Illinois at Chicago Circle*
JANE W. TORREY, *Connecticut College*
CYNTHIA TURNURE, *University of Minnesota*
BENTON J. UNDERWOOD, *Northwestern University*
JAMES F. VOSS, *University of Pittsburgh*
WILLIAM C. WARD, *Educational Testing Service*

* Reviewed more than one manuscript.

Journal of Educational Psychology
1974, Vol. 66, No. 1, 152-156

CONSTRUCT VALIDITY OF TEST ITEMS MEASURING ACQUISITION OF INFORMATION FROM LINE GRAPHS

JAY R. PRICE,[1] VICTOR R. MARTUZA, AND JAMES H. CROUSE

University of Delaware

Research on the effectiveness of graphical displays for information acquisition and retention lacks a system for classifying graph information and generating test items to assess learning. The purpose of this study was to validate a system based on two types of information and three types of informational units. Results of an analysis of variance indicated differences in learning predictable from the classification system; however, a multitrait-multimethod matrix analysis failed to provide evidence of trait validity for the system's informational constructs. In light of these results, a graph information-processing strategy was proposed in which subjects utilize data point information.

The present study deals with the acquisition and retention of quantitative information from a line graph stimulus. While the acquisition of quantitative information from graphical displays is an important component of school learning, the processes involved in such situations have been studied only infrequently, (cf. Schutz, 1961; Washburne, 1927). The present study is particularly concerned with three aspects of learning from a line graph stimulus: (*a*) the nature of the informational unit(s) processed by subjects instructed to learn the information in the graph, (*b*) the relationship between the number of informational units upon which a test item is based and the accuracy of subject performance on that item, and (*c*) the relationship between study time and acquisition of information from the graph.

In attempting to measure the acquisition of information from a line graph stimulus, the first question that arises concerns the nature of the informational units processed by the subject. A logical distinction exists between point and slope information. In a line graph, a unit of point information is the value of the dependent variable associated with a specific level of the independent variable; a unit of slope information is the change in value of the dependent variable per unit change in the independent variable associated with a specific, contiguous set of independent variable levels. The question of immediate interest is whether this logical distinction is a meaningful psychological distinction; that is, when instructed to learn the information in a line graph, do subjects encode point and/or slope information? If subjects do, in fact, store point and slope information independently, then point and slope information can be viewed as informational constructs in much the same way that personality constructs are viewed; thus, it should be possible to validate items measuring these informational constructs by means of multitrait-multimethod methodology (Campbell & Fiske, 1959).

The second question of interest concerns the relationship between the number of informational units required for correct performance on items at recall and accuracy of subject performance on these items. Studies by Schutz (1961) and Washburne (1927) are tangentially related to this question, but because of differences in procedure, task instructions, and type of item presentation format, the studies do not lead directly to expectations for the present experiment. However, it would seem that the greater the number of informational units required by

[1] Requests for reprints should be sent to Jay R. Price, College of Education, University of Delaware, Newark, Delaware 19711.

an item at recall, regardless of the type of unit involved, the poorer performance should be on the item.

The third question of interest concerns the effects of study time on information acquisition. The purpose here was to extend the research on study time into the area of learning quantitative information from graphical materials. It was expected, as most studies have shown, that increased study time would result in greater acquisition. Of greatest interest, however, were the possible interactions of study time with the type of informational units and with the number of informational units that were required for successful performance on the test items at recall.

Method

Subjects

Thirty-six undergraduate education student volunteers served as subjects in this experiment.

Materials

A multiple-line graph was constructed in which the average value per share of stock for each of three fictitious companies was plotted for each of five successive years. Each of the three lines (one per company) was generated randomly, subject to the following constraints: (*a*) One line would show an increasing trend; (*b*) the second line would show a decreasing trend; and (*c*) the third would show random fluctuations. To generate the data points for the first two of these lines, the data point values were randomly sampled from the following five strings of digits: 0–5, 1–6, 2–7, 3–8, and 4–9. For the increasing trend line, the first digit was randomly selected from the 0–5 interval. The next four digits were randomly selected from the four succeeding digit strings. For the decreasing trend line, the first digit was randomly selected from the 4–9 interval. The next four were randomly selected from the remaining intervals in sequence. The five values for the third line were randomly selected from the 0–9 range subject to the restriction that there would be exactly one intersection or crossover of lines in the left, center, and right thirds of the graph.

The criterion test consisted of six subtests of eight propositions each. Three subtests were based on point information, the rest on slope information. Within each information type, the three subtests were based on a single unit of information, two units arranged vertically (i.e., the price of stock for two companies during the same year) and two units within the same line (i.e., the price of a single company's stock for two separate years). Following the lead of Anderson (1972), Bormuth (1970), and Cronbach (1971), basic sentence frames were formed for each item type (see Table 1), and rules were established to generate the items in each cell.

For example, the rules for the point items based on a single unit of information are listed below:

1. Company names for the eight items were selected randomly with the restriction that each company name was used at least twice and no more than three times.
2. The year values for the eight items were chosen randomly with the restriction that each year value was used at least once and no more than twice.
3. The comparative (greater than – less than) was assigned randomly to the items so that each appeared in four items of the subtest.
4. Within the four items containing the "greater than" comparative, the truth value was randomly assigned such that two propositions would be true and two would be false. The same procedure was used for the four "less than" comparative items.
5. For each item, the set of stock values that would satisfy the truth value for that item was determined, and one element of the set was randomly selected for inclusion in the item.

It is apparent from the above rules that items

TABLE 1

Test Item Frames: Information Type and Number of Informational Units

Information type	Single unit	Two units within occasion	Two units within group
Point	The average value per share of (*name*) stock was (±*comparative*) than (*value*) during the year (*year*).	The average value per share of (*name*) stock was (±*comparative*) than the average value of (*name*) stock during the year (*year*).	The average value per share of (*name*) stock was (±*comparative*) in (*year*) than in (*year*).
Slope	The average value per share of (*name*) stock (±*verb*) during the period (*years*).	The average value per share of (*name*) stock changed (±*comparative*) rapidly than the value of (*name*) stock during the period (*years*).	The average value per share of (*name*) stock changed (±*comparative*) rapidly during the period (*years*) than during the period (*years*).

within each subtest were balanced for wording of comparatives (e.g., greater than – less than, more rapidly – less rapidly, increased – decreased) and truth value. With respect to wording of comparatives, a number of researchers (e.g., Clark, 1970, Trabasso, 1971) have shown that positive and negative wording of test items impose different information-processing requirements on subjects with resulting differences in performance levels. These results as well as those on acquiescent responding suggested that items should be balanced for comparative wording and truth value so that comparisons of interest would not be differentially contaminated by differences in responding.

Analogous procedures were used for generating each of the five remaining item types. The items were then randomly ordered over the test as a whole, subject to constraints necessary for guaranteeing that the distribution of the various item characteristics described above would be even across the test as a whole.

The graph and test items were reproduced on standard 8½ × 11 inch sheets of paper and bound in a seven-page test booklet. A cover sheet for subject identification was followed by the graph. A blank sheet followed the graph and separated it from the three pages of test propositions to prevent the subjects from seeing the graph at test time. A final cover sheet completed the test booklet.

Procedure

The subjects were randomly assigned in equal numbers to the 2- and 8-minute study time conditions. Following distribution of the materials, instructions were read to the subjects that (*a*) indicated the purpose of the study, (*b*) specified both the study time and test time limits, (*c*) informed them that the graph could not be used as a reference once the prescribed study time had elapsed, and (*d*) instructed them to answer all items. Subjects were told they had up to 40 minutes to complete the test items. As it turned out, no one required more than 25 minutes to complete the test items.

Results

The number of correct responses per item type was determined for each subject. These data were then analyzed as a one-between, three-within factorial analysis of variance. The between factor was study time and the within factors were information type, number of informational units, and wording of logical opposite pairs. Table 2 contains the means and standard deviations for this analysis.

All four main effects were significant, while none of the interactions were significant. The mean score in the eight-minute study condition was higher than was the mean in the two-minute condition (F = 10.90, df = 1/34, p <.01). The mean score on point information items was significantly higher than was the mean on slope information items (F = 6.18, conservative df = 1/34, p <.02). Scheffé tests on the three information unit means indicated that the mean of single-unit items was higher than were the weighted means of the two unit-within-occasion and two unit-within-group items (p < .01); however, the means of the latter two item types were not significantly different from each other (p > .05). The mean performance on items stated positively (greater than, increase, more rapidly) was significantly higher than was the mean performance on items stated negatively (F = 6.16, conservative df = 1/34, p < .02).

To assess the relationship between performance on information types and number of data points required to answer an item

TABLE 2

Means and Standard Deviations for the Study Time × Information Type × Units of Information × Wording Interaction

Study condition	Point						Slope					
	One unit		Two within group		Two within occasion		One unit		Two within group		Two within occasion	
	+	−	+	−	+	−	+	−	+	−	+	−
Two minute												
$\bar{X}$	3.22	2.89	3.06	2.61	3.11	2.72	3.17	2.94	2.56	2.67	2.44	2.28
SD	1.08	.99	.71	1.11	.93	.87	1.01	.91	.89	1.00	.95	1.15
Eight minute												
$\bar{X}$	3.89	3.33	3.11	3.06	3.56	3.56	3.28	3.44	3.28	2.89	3.44	3.11
SD	.31	.82	.93	1.03	.60	.60	.81	.60	.86	1.15	.84	.93

TABLE 3

COMPARISONS AMONG SUBTEST MEANS

No. of data points	Information type	No. of units	$\bar{X}$	Subtest	Subtest 6	5	2	4	3	1
4	Slope	2 within occasion	5.639	6		.55	.278	.778	.833	1.028*
3	Slope	2 within group	5.694	5			.223	.723	.778	.973*
2	Point	2 within group	5.917	2				.500	.555	.750
2	Slope	1	6,417	4					.055	.250
2	Point	2 within occasion	6.472	3						.195
1	Point	1	6.667	1						

* $p = .01$.

successfully, the six subtest means (Information Type × Number of Units) were analyzed as a one-between, one-within analysis of variance (Time × Subtest). The two main effects were significant; the interaction was not. The number of data points and subtest means as well as the significant comparisons by the Newman-Keuls procedure are contained in Table 3. This analysis indicated that only the mean of the point-single-unit test differed significantly from the means of the slope-within-occasion and slope-within-group tests.

In order to assess possible effects of response sets, the data were reanalyzed with study time, logical opposite pairs, and truth value as the independent variables. The only significant results were those main effects associated with study time and logical opposite wording. The fact that all interaction effects were nonsignificant seemed to rule out acquiescence as a possible explanation for the results obtained in the initial analysis discussed above.

TABLE 4

MULTITRAIT (POINT AND SLOPE INFORMATION TYPE) – MULTIMETHOD (NUMBER OF INFORMATIONAL UNITS) MATRIX

Measure	1		2		3	
	Point	Slope	Point	Slope	Point	Slope
1. Single unit						
Point	(.42)[b]					
Slope	.07	(.47)				
2. Within group						
Point	.93	.80	(.48)			
Slope	1.00	.62	1.00[a]	(.37)		
3. Within occasion						
Point	.15	1.00[a]	.91	.88	(.48)	
Slope	.15	.76	.20	.21	.31	(.57)

[a] Actual corrected values greater than 1.

[b] Numbers within parentheses represent the reliability estimate according to the Kuder Richardson Formula 20.

Table 4 contains the multitrait-multimethod matrix with number of informational units representing the methods, and point and slope information being the possible constructs. Correlation coefficients appearing in the table have been corrected for attenuation. The overall pattern of coefficients in the matrix does not support our hypothesis that the point and slope items included in this criterion test measure two distinct informational constructs.

DISCUSSION

The results of the initial analysis indicated significant main effects for study time, wording, number of informational units, and informational types. The effect of informational types suggested that the point–slope dichotomy was a meaningful distinction; however, the multitrait-multimethod matrix failed to support this distinction: Performance on the various point and slope subtests predicted performance on subtests both within and between these two informational constructs.

An explanation for the disparate results of these two analyses may lie in the kind of information subjects encoded and/or retrieved under the experimental instructions and conditions of this study. It is possible that subjects did not use slope information

as defined in this study but instead used only data point information. To answer slope items, subjects recalled point information and then constructed slope information from the recalled points. The reasoning that follows supports this conclusion.

Slope items are apparently more difficult than point items. If slope performance is a function of a subject's recall of data points, then an increase in the number of data points needed for successful performance should be accompanied by a decrease in performance level. From Table 3, it is apparent that this inverse relationship exists; subjects' scores tend to decrease as the number of data points increases.

Consequently, it appears that the amount of data point information may be a more important factor than informational type in determining a subject's performance level given the proposed information-processing strategy. However, the present findings do not rule out the possibility that under other experimental instructions and conditions, subjects might encode slope information. If this were the case, then the present multitrait-multimethod methodology would be suitable for providing evidence of the encoding of slope information and the validity of the slope informational construct.

REFERENCES

ANDERSON, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research,* 1972, **42,** 145–170.

BORMUTH, J. *On the theory of achievement test items.* Chicago: University of Chicago Press, 1970.

CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, **56,** 81–105.

CLARK, H. H. How we understand negation. Paper presented at the Workshop on Cognitive Organization and Psychological Processes, Huntington Beach, California, August 1970.

CRONBACH, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement.* Washington, D.C.: American Council on Education, 1971.

SCHUTZ, H. G. An evaluation of formats for graphic trend displays—Experiment II. *Human Factors,* 1961, **3,** 99–107.

TRABASSO, T., ROLLINS, H., & SHAUGHNESSY, E. Storage and verification stages in processing concepts. *Cognitive Psychology,* 1971, **2,** 239–289.

WASHBURNE, J. N. An experimental study of various graphic, tabular, and textural methods of presenting quantitative material. *Journal of Educational Psychology,* 1927, **18,** 361–376, 465–476.

(Received February 28, 1973)

Books Geared to Educational Psychology Today

JANICE T. GIBSON, of the University of Pittsburgh on

Educational Psychology, 2nd Ed., 1972

Provides programmed instruction for in-training and in-service teachers at all levels. Presents viewpoints of a psychologist experienced in research methods of psychology and in the application of these methods to classroom problems.

This edition treats differences between people in abilities, aptitudes, and attitudes, and deals with problems involved in measuring these differences. Discusses theories of learning instruction together with evolving instructional procedures—from computer assisted instruction to informal procedures of the open classroom.

Programmed instruction provides the reader with small steps, active participation, and immediate knowledge of results.

Accompanying Instructor's Manual includes cross-referenced chart correlating chapters in this text with other chapters in major educational psychology books. 1972, 350 pp., spiral-bound $7.15 (Formerly published by Appleton, Century, Crofts.)

For further information, write: Robert Jordan, Dept. J-846

JOHN P. DE CECCO, of California State University, San Francisco; and WILLIAM R. CRAWFORD, of PRIMEX Project, University of California, Los Angeles on

The Psychology of Learning and Instruction: 2nd Ed., 1974

EDUCATIONAL PSYCHOLOGY—the successful first edition has been revised and updated to give a broader appeal to the student of the '70s. The new edition blends humanistic psychology with traditional, but retains its empirically-based educational practice orientation. Suitable for the basic course in educational psychology generally required in teacher-preparation programs.

Includes new chapters on intellectual development of children using Jean Piaget's theory and uses modern linguistic and psycho-linguistic theory in the language development of children. Also offers new updated research and reference material, and improves the teaching practice of all four components of the basic teaching model. February 1974, 608 pp., cloth $10.95

Also available—new and updated Student Guide, plus new and updated Teacher's Manual both by Anne F. Terrill, University of Pittsburgh, Pennsylvania.

Prentice-Hall, Englewood Cliffs, N.J.

Sprinthall & Sprinthall isn't really a book.

It's a wide-angle lens.



Richard C. Sprinthall and Norman A. Sprinthall designed this remarkable lens. They call it EDUCATIONAL PSYCHOLOGY: A DEVELOPMENTAL APPROACH.

It's a lens that comprehensively surveys the ed psych landscape. Where ed psych came from. Where it is. Where it can go. It puts the pieces together and makes them workable.

The focal point is the student. His physical, cognitive, emotional and moral development. From birth through adolescence.

In the background and interwoven throughout the scene are major figures capsuled in a series of vignettes. They summarize a textual maincurrent, pull together ideas and methods via a single personality. Freud, Bruner, Skinner, Piaget and others.

It's a visual experience. Made cogent by the Sprinthalls' writing style. Lucid. Comfortable. Often entertaining.

One who scrutinized it had this to say: "If the book were published tomorrow, I would use it the very next time I taught educational psychology. It is the best all-round book I know of."

We agree, of course. Although we really think it's more like a lens.

College Division
ADDISON-WESLEY PUBLISHING COMPANY, INC.
Reading, Massachusetts 01867

Allyn and Bacon, Inc.

New 1974
Contemporary Issues in Educational Psychology,
Second Edition. Edited by Harvey F. Clarizio, Robert C. Craig, and William A. Mehrens, all of Michigan State University. 1974, 6⅛x9¼, est. 756 pp.

A Sympathetic Understanding of the Child: Birth to Sixteen
By David Elkind, University of Rochester. 1974, 5¼x8¼, paperbound, est. 216 pp.

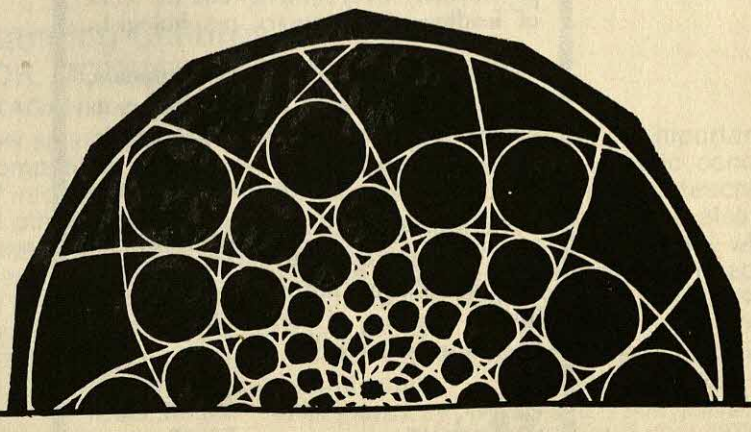
Teachers and Learners: The Interactive Process of Education, Second Edition. By Alfred H. Gorman, Montclair State College. 1974, 5½x8½, paperbound, est. 228 pp.

Teaching/Discipline: A Positive Approach for Educational Development, Second Edition. By Charles H. Madsen, Jr. and Clifford K. Madsen, both of Florida State University. 1974, 5⅜x8⅜, paperbound, est. 150 pp.

Educational Psychology, Fourth Edition
By James M. Sawrey and Charles W. Telford, both of San Jose State College. 1973, 7x9¼, paperbound, 706 pp.

Human Dynamics in Education and Psychology: Selected Readings, Second Edition. Edited by Don E. Hamachek, Michigan State University. 1972, 5⅝x8⅜, paperbound, 667 pp.

Allyn and Bacon, Inc., College Division, Dept. 893
470 Atlantic Avenue, Boston, MA 02210



"An intellectual adventure of the first order."*

JEAN PIAGET

The Man and His Ideas

Richard I. Evans

A DIALOGUE WITH PIAGET
plus illuminating articles by
Jean Piaget, David Elkind and others

IN THIS DIALOGUE

Intelligence Testing

Applying His Ideas to Education

Comparisons with Freud, Skinner and Erikson

The Intellectual Development of the Child

Through its innovative dialogue format, comprising a series of structured but informal interviews with Jean Piaget, this book provides both the student and general reader with a clear, undistorted understanding of Piaget's theories. Piaget also gives his views on Gestalt psychology, psychoanalysis, cybernetics and the work of leading contemporary psychologists.

"Throughout the man himself appears, speaking eloquently, a great teacher as he undoubtedly is...informative, inspiring and exemplary."
—**Los Angeles Times*

"The bibliography is outstanding; and the unexpected and most welcome bonus of Piaget's autobiography—much in the manner of Henry Adams—has been sorely needed."
—*Library Journal*



$8.95, cloth
$2.95, paper

Key textbooks for learning about children and adolescents

The Adolescent Years

Second Edition

WILLIAM W. WATTENBERG, *Wayne State University*

The Second Edition of this successful textbook for courses in adolescent psychology continues to offer a thorough, well-documented analysis of all phases of adolescent development. Written in an informal, highly readable style, the book presents adolescence as a developmental stage, describing how childhood experiences alter adolescence and showing in general what happens during pre-adolescence, adolescence, and the years of near-adulthood. The book is effectively organized into four major parts or cycles: Part A presents an overview of general adolescent characteristics; Part B deals with influences than can modify the general pattern; and Part C is concerned with behavioral settings in the adolescent life—among them, the home, the school, and the community. Finally, Part D takes up the major problems in adolescent life, including discussions of money, sex, delinquency, drugs, and careers. Throughout the book the author illustrates the text material with case studies, often presented in the form of vignettes.

458 pages. $10.95

Readings in Child Behavior and Development

Third Edition

Edited by CELIA STENDLER LAVATELLI, *University of Illinois,*
and FAITH STENDLER, *Massachusetts Mental Health Center*

Widely adopted and highly praised for its content and currency, the new Third Edition of *Readings in Child Behavior and Development* presents 4€ important articles reflecting the most recent and significant research anc thinking in the field. Each of the six parts has a detailed introduction tha surveys what is important in the various domains of child development. The first part, "The Grand Systems," provides a solid theoretical foundation fo approaching the individual papers and includes articles by Erikson, Skinner Piaget, Werner, Anna Freud, and Hess. The book then goes on to stress such major contemporary concerns as cognitive development, infancy, the pre-school child, cross-cultural differences, and the disadvantaged in classic papers by noted psychologists. Paperbound. 529 pages. $6.95

Understanding Children

BEHAVIOR, MOTIVES, AND THOUGHT

JEROME KAGAN, *Harvard University*

"This small book is a gem. It accomplishes two major, important, and often incompatible objectives. It develops significant scientific concepts, ideas, and information for non-scientific audiences—in this case teachers, parents, and other adults interested in and puzzled by the young. It also provides an informal but coherent outline of cognitive theory, a theory which, in this instance, incorporates motivation and emotion as well as concept attainment and thinking. . . . Here is a psychological frame of reference which humanistic educators, and many others, can use to array their ideas, understandings, and energies as they engage in pressing problems of parenting, guiding, teaching, and counseling children and youth." Paperbound. 153 pages. $3.50

From a review by Dale B. Harris, Pennsylvania State University in *Contemporary Psychology,* 1972, Vol. 17, No. 6



HARCOURT BRACE JOVANOVICH, INC.

New York / Chicago / San Francisco / Atlanta

intext

CHANGING CLASSROOM BEHAVIOR, 2ND ED.

Merle L. Meacham, University of Washington and Allen E. Wiesen

The authors advocate "objective teaching," a method that utilizes reinforcement, extinction, imitative learning and programming—and show how their techniques can help the child by giving him skills that give *him* some control over his environment and his behavior. Many practical illustrations show how to use these and related techniques in class.

February, 1974 288 pages paper, about $4.95

STATISTICAL METHODS FOR BEHAVIORAL SCIENCE

Billy L. Turney, Tyler State College and George P. Robb, North Texas State University

A thorough introduction to the concepts and procedures for both descriptive and inferential statistics, including problems that supply feedback to facilitate problem-solving. Exercises following each topic have their answers printed on the reverse page, so students can check themselves as they read.

1973 214 pages paper, $6.95

THE ADOLESCENT: Case Studies for Analysis

Nancy C. Ralston and G. Patience Thomas, both of the University of Cincinnati

After an introduction to the case study method and a summary of basic psychological theories about learning, development, and behavior, the authors present nine case studies on the varied problems of adolescence—including one on suicide and one on minority groups.

March, 1974 208 pages paper, about $4.95

THE CHILD: Case Studies for Analysis

Nancy C. Ralston and G. Patience Thomas, both of the University of Cincinnati

Eighteen case studies with interpretive questions that will help students develop a broad understanding of some of the typical behavior patterns of childhood, along with insight into the causes and meaning of different kinds of behavior. A sample case study outline is included.

1972 200 pages paper, $3.95

iep

INTEXT EDUCATIONAL PUBLISHERS · 257 Park Avenue South · New York, N.Y. 10010

STOP—LOOK

BUILDING SKILLS FOR COMPETENCY-BASED TEACHING

Leo D. Leonard and Robert T. Utz

Here is the first book to develop the basic skills for building a competency-based curriculum by using its own methodology as a model. It begins with an explanation of the characteristics of competency-based education and a pretest covering the basic skills. For each type of skill—developing student self-discipline, applying learning concepts, formulating educational objectives, developing a curriculum, and evaluation—the relevant behavioral objectives are identified and discussed using numerous sample problems, and the reader is guided in evaluating himself. Finally, the book asks the student to synthesize the basic skills and to develop a competency-based curriculum for his own particular area of teaching. Charts; glossary; chapter bibliographies. February 1974. Tentative: 248 pages; $5.95/paper.

COGNITIVE PROCESSES IN EDUCATION:

A Psychological Preparation for Teaching and Curriculum Development

Sylvia Farnham-Diggory

Devoted exclusively to the study of psychology as it applies directly to the methods and problems of American education, this book explains why such recent developments as the new math and the new social studies are psychologically more meaningful than traditional programs. Relevant research findings are compiled and used to formulate a psychological rationale for better education. Case histories illustrate the author's development of such themes as impulsivity, dependence, prejudice, and rigidity in the classroom. Actual curriculum examples and materials illustrate explanations. Published. 630 pages; $12.50. *Instructor's Manual. Study Guide*: $1.95.

LEARNING THEORIES FOR TEACHERS,

Second Edition

Morris L. Bigge

A balanced presentation of early and contemporary learning theories and their implications for teaching, this book avoids oversimplification of the basic tenets and describes similarities and differences among theories to guide the student in critically constructing and evaluating his own outlook on the nature of learning. Discussion of the strengths of each theory is designed to confront the reader with the need to choose among competing psychological points of view. Weaknesses are left for the reader to discern. Published. 358 pages; $5.35/paper.

Harper & Row

offers these texts and many

more→

FOR A VARIETY

PSYCHOLOGICAL CONCEPTS IN THE CLASSROOM

Richard H. Coop and Kinnard P. White

Organized around a set of scenarios, this text describes three types of schools in American society—an inner city school, a small town-rural school, and a school in a wealthy suburban area—and presents case studies drawn from each. Eight specific conceptual areas are treated: power and influence in the classroom, self-concept, teaching expectancy, behavior modification, locus of control, motivation, intelligence, and information processing. Each chapter presents an overview of the research in a given area and relates the evidence directly to the school scenarios or the individual case studies. Problems and study questions; illustrations; and bibliography. February 1974. Tentative: 320 pages; $4.95/paper.

LOOKING IN CLASSROOMS

Thomas L. Good and Jere E. Brophy

The need for teachers to become more aware of their classroom behavior is the central theme of this book. It surveys recent advances in educational research and gives detailed advice about effective teaching. Among the subjects covered: Teachers' attitudes and expectations, the teacher as a behavioral model, classroom organization and management, ability grouping, and peer tutoring. Published. 397 pages; $4.95/paper.

MAINTAINING SANITY IN THE CLASSROOM: Illustrated Teaching Techniques

Rudolf Dreikurs, Bernice Bronia Grunwald, and Floy C. Pepper

Based on the psychology of Alfred Adler and heavily case-oriented, this text applies techniques of motivation modification to classroom problems. The authors systematically utilize their own case material to illustrate teaching problems and how to deal with them. Published. 338 pages; $5.35/paper.

Harper & Row

10 East 53d Street, New York 10022

OF EDUCATIONAL PSYCHOLOGY COURSES

STATISTICS FOR EDUCATION: With Data Processing

David White

This introductory statistics text gives education majors actual experience in using statistical techniques to solve educational problems. The inclusion of raw educational data from two school districts familiarizes students with the common characteristics of such data and provides them with experience in interpreting evidence for or against meaningful propositions. The book features material on the use of high-speed data processing facilities, sampling methods, and decision-making. Solutions to problems and exercises are presented. Just published. 382 pages; $10.95.

PROBLEM SITUATIONS IN TEACHING

Gordon E. Greenwood, Thomas L. Good, and Betty L. Siegel

This text provides a decision-making view of teaching through 20 representative case studies, developed from data on over 330 problem cases. A strong, detailed introduction suggests a strategy for attacking teaching problems. The cases, each reported behaviorally and left unresolved, allow for original student analysis and solutions. Stimulus questions following each case serve to provoke and guide discussion. *Instructor's Manual.* Published. 162 pages; $3.95/paper.

PSYCHOLOGICAL FOUNDATIONS OF EDUCATION, Second Edition

Morris L. Bigge and Maurice P. Hunt

This problem-oriented text takes a semihistorical and comparative approach in discussing how children develop through adolescence, how they learn, the relationship between development and learning, and how knowledge of this relationship promotes effective teaching. Cognitive-field theory is contrasted with S-R associationistic and other views. The text differentiates among schools of thought in psychology to offer a choice of teaching methods and to provide a basis for developing cognition and intellectual procedures in the classroom. Classroom application of psychological theory is emphasized. Schematic line drawings; case studies; references; annotated bibliographies. *Instructor's Manual.* Published. 603 pages; $11.50.

MORE TEXTS

CHILD DEVELOPMENT AND PERSONALITY, Fourth Edition

Paul Henry Mussen, John Janeway Conger, and Jerome Kagan

736 pp. Tentative: $11.95; March, 1974. For use with the text: New Instructor's Manual; New Study Guide by Fay-Tyler Norton; New PSI Study Guide by Ray DeV. Peters; New PSI Instructor's Manual by Ray DeV. Peters.

BASIC STATISTICAL METHODS, Fourth Edition

N. M. Downie and Robert W. Heath

355 pp.; $10.95 (tentative); March, 1974. Study Guidebook.

CHILD DEVELOPMENT: The Human, Cultural, and Educational Context

W. H. O. Schmidt

181 pp.; Paper: $3.50; July, 1973.

ADOLESCENCE AND YOUTH: Psychological Development in a Changing World

John Janeway Conger

573 pp.; $10.95; June, 1973. Instructor's Manual.

ESSENTIALS OF PSYCHOLOGICAL TESTING, Third Edition

Lee J. Cronbach

752 pp.; $11.75; 1970.

EXERCISES IN PSYCHOLOGICAL TESTING

Eugene R. Oetting and George C. Thornton III

229 pp.; Paper: $5.25; 1968.

UNDERSTANDING SCHOOL LEARNING: A New Look at Educational Psychology

Michael J. A. Howe

299 pp.; Paper: $3.50; 1972.

from Harper & Row 1817

Send for our 1974 educational psychology catalog (CT08)
Harper & Row/10 East 53d Street, New York 10022

The only thing better than Biehler

The Second Edition of Robert Biehler's **Psychology Applied to Teaching** is the only introduction to educational psychology that could be better than the First Edition.

Because the new Biehler has all the best features of the old, and then some.

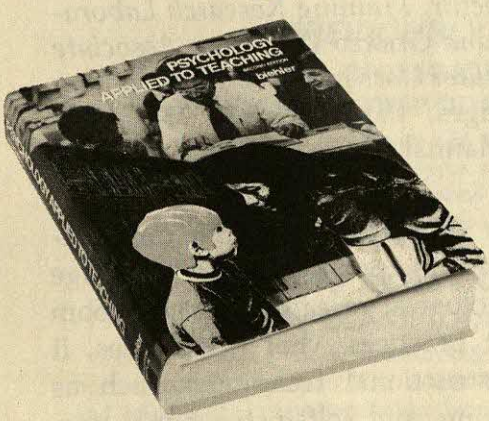
The Second Edition is as readable and as teachable as ever, and Biehler preserves the definitive approach of the First Edition — applying psychology to actual classroom situations.

But the new Biehler has been revised and updated, too.

The text now covers performance contracting, accountability, free schools, open education, inequality of educational opportunity, and the Nature-Nurture Debate. Biehler also expands coverage of the theories of Bruner, Piaget, Erikson, and Maslow.

The Second Edition of **Psychology Applied to Teaching.** It's the only thing that could be better than the First Edition of Biehler.

And nothing is better than a better Biehler.



Psychology Applied to Teaching Second Edition
Robert Biehler
California State University at Chico
795 pages
with SELECTED READINGS (1972), and new Study Guide and Instructor's Manual / January 1974

is a better Biehler

Houghton Mifflin Boston 02107 / Atlanta 30324 / Dallas 75235
Geneva, Ill. 60134 / Hopewell, N.J. 08525 / Palo Alto 94304



EDUCATIONAL PSYCHOLOGY:
focus on the learner

Lita Linzer Schwartz, The Pennsylvania State University, Ogontz Campus

COMPREHENSIVE: The text includes the topics traditionally covered in beginning educational psychology courses; the author's ability to weave interesting examples into the text heightens motivation and makes the principles more concrete.

FOCUS SEGMENTS: Focus segments clearly illustrate case studies, experimental research, and practical applications of theoretical principles.

TIMELY: Innovative topics such as educational accountability and the British approach to primary education are included.

TEACHING AIDS: Outlines, glossaries, bibliographies, and charts clarify the information contained in the chapters. An instructor's manual is available upon adoption. **1972, 7½ x 9¼, 604 pp.**

HOLBROOK PRESS, INC. Dept. A93, 470 Atlantic Ave., Boston, MA 02210 (a subsidiary of Allyn and Bacon, Inc.)

EDUCATIONAL PSYCHOLOGY: *The Science of Instruction and Learning*

By Richard C. Anderson, *Director, Training Research Laboratory, University of Illinois, and* Gerald W. Faust, *Associate Director, Institute for Computer Uses in Education, Brigham Young University.* 528 pages. Paperback. $7.95. Study Guide, $2.95. Instructor's Manual.

Everyone who is seriously concerned with the improvement of instruction should read this practical effort to bridge the gap between recent psychological research and classroom applications. A textbook that practices what it preaches, it incorporates a variety of instructional techniques such as behavioral objectives, self-testing, and self-instructional programs. This experiential dimension, the authors believe, is what is most likely to increase the probability of good teaching.

DODD, MEAD & COMPANY,

79 Madison Avenue, New York, New York 10016

From Macmillan—

The Current Topics in Classroom Instruction Series

All by Norman E. Gronlund, University of Illinois

COMING IN 1974

INDIVIDUALIZING CLASSROOM INSTRUCTION

Surveys many of the comprehensive systems of individualized instruction currently in use across the country, describes some of the common classroom procedures and techniques, and provides sources of information on other programs currently being developed.

1974 approx. 64 pages paper prob. $1.50

DETERMINING ACCOUNTABILITY FOR CLASSROOM INSTRUCTION

A practical guide to understanding accountability, its principles and practices, that enables in-service and prospective teachers to more effectively participate in school accountability programs.

1974 approx. 64 pages paper prob. $1.50

IMPROVING MARKING AND REPORTING IN CLASSROOM INSTRUCTION

1974 approx. 64 pages paper prob. $1.50

PREVIOUSLY PUBLISHED

PREPARING CRITERION-REFERENCED TESTS FOR CLASSROOM INSTRUCTION

Shows teachers how to prepare and use criterion-referenced tests. Such tests are designed to interpret specific types of learning achievement and to yield precise descriptions of the student's success in learning.

1973 55 pages paper $1.50

STATING BEHAVIORAL OBJECTIVES FOR CLASSROOM INSTRUCTION

Trains students to state objectives for classroom instruction in terms of expected behavior outcomes. The author describes the methods used in identifying and defining the instructional objectives as learning outcomes.

1970 58 pages paper $1.65

MACMILLAN PUBLISHING CO., INC.

100A Brown Street, Riverside, New Jersey 08075

Important Macmillan

EDUCATIONAL PSYCHOLOGY

A Scientific Foundation for Educational Practice

Robert M.W. Travers, Western Michigan University

Following an introductory discussion of the need for scientific information to solve problems in education, *Educational Psychology* follows the development of learning from birth through the school years. It features:

- Classroom examples that illustrate the scientific principles underlying learning and human development, achievement and motivation, retention, social development, and pupil assessment.
- Consideration of such special problems as those of transfer, evaluation, and appraisal of learner potential.
- Detailed chapter summaries to facilitate study and review.
- Many examples of teacher-pupil interactions.
- An accompanying teacher's manual that contains over 300 multiple-choice test items.

1973 448 pages $9.95

MEASUREMENT AND EVALUATION IN TEACHING

Second Edition

Norman E. Gronlund, University of Illinois

This book introduces teachers and prospective teachers to the principles and procedures of measurement and evaluation that are essential to effective teaching. Among the revisions in the second edition are two new chapters—on standardized achievement testing and standardized aptitude testing. However, primary emphasis is still on teacher-made tests. Other changes include:

- Revised, updated material on validity and reliability.
- New material on how to state instructional objectives in behavioral terms.
- New information on the role of evaluation in instruction, curriculum development, and programmed learning.
- Condensed descriptions of the taxonomies of educational objectives, with illustrative objectives and lists of behavioral terms.

1971 545 pages $9.95

Educational Psychology Books

BEHAVIOR MODIFICATION IN EDUCATION

Donald L. MacMillan, University of California, Riverside

This new text presents the principles and practices of behavior in an educational context to facilitate classroom application. Professor MacMillan provides concrete examples of translating principles into action while opening new vistas for imaginative teaching. General principles of behavior modification are discussed and related to general learning theory and teaching. Professor MacMillan feels that unless the general principles are well understood, prospective teachers are tempted to adopt other educators' practical programs *in toto* without considering their appropriateness to their own teaching situations. In addition, this thorough grounding encourages teachers to experiment with different applications of specific principles.

Coverage includes behavior acquisition, maintenance and extinction, reinforcement scheduling, and target behavior selection.

1973 238 pages paper $4.95

PRACTICAL RESEARCH

Planning and Design

Paul D. Leedy, The American University

Practical Research is a unique, how-to guidebook to the practical aspects of research. Step by step, the book tells not only what research is, but what the tools of research are and how they are employed by the researcher; how a research problem is selected; how the procedures relating to the problem are carried out; and how the written report is presented for maximum effectiveness. The book is clear and to the point. It spells out much that has been taken for granted in other research texts. The author includes many unique learning aids to enhance and further clarify the textual material. They include a sample dissertation proposal in the appendix that has marginal notes indicating the proper form of presentation; and a practicum in the appendix that allows the student to again experience each phase of the textual coverage by actually planning and presenting a research proposal.

1974 approx. 256 pages paper prob. $4.95

MACMILLAN PUBLISHING CO., INC.
100A Brown Street, Riverside, New Jersey 08075

DICTIONARIES

WEBSTER

Library Size 1973 edition, brand new, still in box.

Cost New $45.00

Will Sell for $15

Deduct 10% on orders of 6 or more

Make Checks Payable to

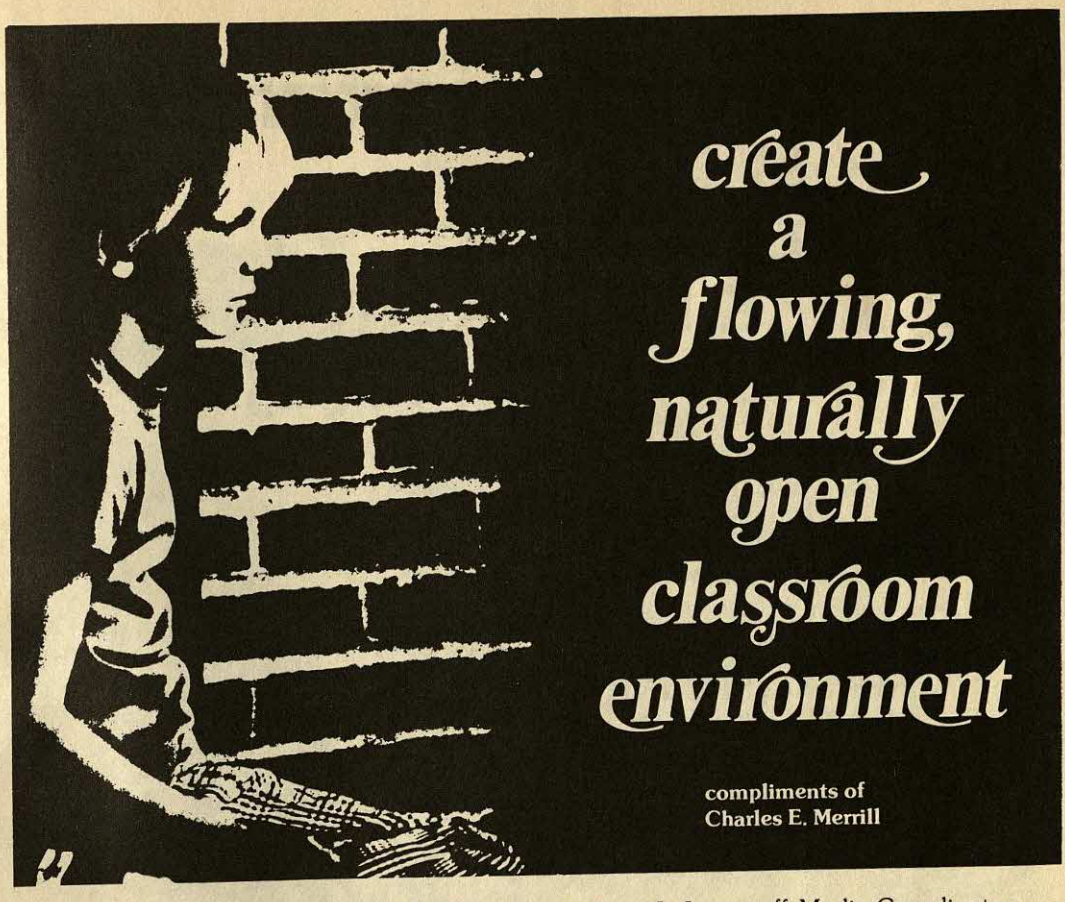
DICTIONARY LIQUIDATION

and mail to

Ontario Text Editions
Dept. D-216
Toronto Dominion Centre
Suite 1400 - 14th Floor
Toronto, Ontario, Canada M5K 1B7

C.O.D. orders enclose 1.00 good will deposit. Pay balance plus C.O.D. shipping on delivery. Be satisfied on inspection or return within 10 days for full refund. No dealers, each volume specifically stamped not for resale.

Please add $1.25 postage and handling



Structuring the Classroom for Success by Anne L. Langstaff, Media Coordinator, Instructional Materials Center for Special Education, University of Southern California; Cara B. Volkmer, Educational Consultant, Periphery; and Marilyn Higgins, Independent Consultant on Media Product Development. This new self-study package is especially designed to assist future elementary teachers in creating, structuring, and maintaining a productive learning environment.

Six sound filmstrips portray concepts that are expanded in the accompanying guidebook. The authors combine principles from the behavior management field with those of open education. Content deals with teacher role, classroom environment, creating activity centers, managing student behavior, individualizing instruction, and evaluating learning outcomes. April, 1974/192 pp./7⅜x9¼"/$4.95 softbound/#8813

Psychology of Classroom Learning: An Inductive Approach by Richard M. Gorman, author of *Discovering Piaget* (Merrill, 1972). February, 1974/488 pp./6⅝x9¾"/$11.95/#8847

Test Construction, 2nd Edition by Dorothy C. Adkins, University of Hawaii. February, 1974/160 pp./6x9"/$3.50 softbound/#8845

For a free examination copy, send your course title, number of students, and name of current text to **BOYD LANE, Merrill, Box 508, Columbus, Ohio 43216.**

Charles E. Merrill Publishing Company

A Bell & Howell Company/Columbus, Ohio

BELL & HOWELL

ISOLATING LEARNING FROM DEVELOPMENT IS LIKE BREAKING A CHILD INTO PIECES.



A child learns and develops as a whole person.
Now there is a textbook built on this simple fact.

In HUMAN DEVELOPMENT AND LEARNING, SECOND EDITION
by Hugh V. Perkins

development is seen as a continuing process. Learning is made practical and easy to understand. They are both presented as integral parts of a whole: the study of human behavior.

Significant features of the Second Edition:
- A new chapter on early infancy and childhood.
- A new chapter on adulthood, including old age.
- Much more attention has been given to culturally disadvantaged and minority group children.
- Expanded section on self-concept, drug use, intelligence and accountability.

Human Development and Learning, Second Edition:
perfect for a two semester psychological foundations course or a comprehensive one semester course. It can also be used for educational psychology or child development or adolescent development courses.

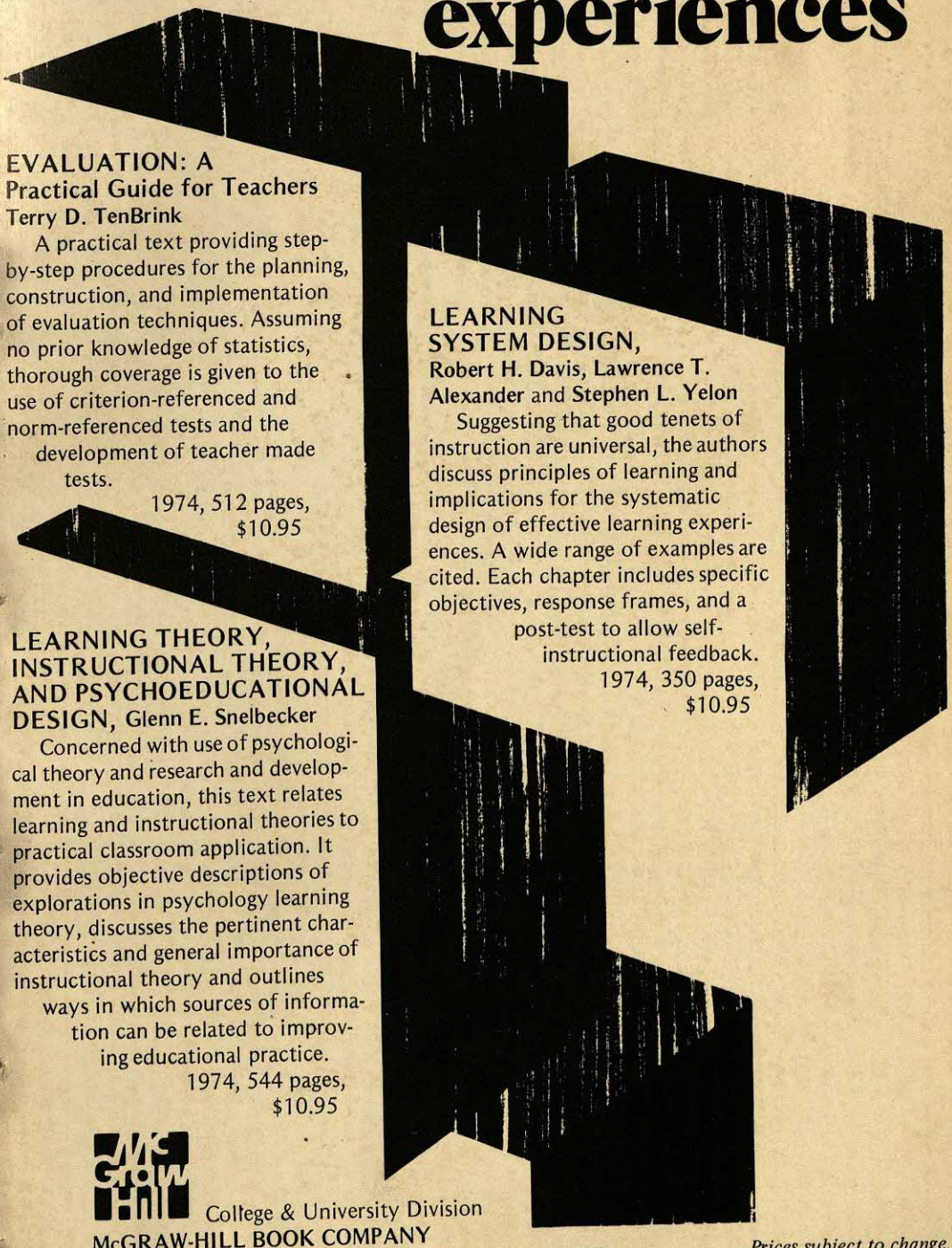
Casebound 6½ x 9¼ price: $12.95 available March 1974 640 pages

For more information write Box JEP-7

WADSWORTH PUBLISHING COMPANY

BELMONT, CALIFORNIA 94002

Build more effective learning experiences



EVALUATION: A Practical Guide for Teachers
Terry D. TenBrink

A practical text providing step-by-step procedures for the planning, construction, and implementation of evaluation techniques. Assuming no prior knowledge of statistics, thorough coverage is given to the use of criterion-referenced and norm-referenced tests and the development of teacher made tests.

1974, 512 pages, $10.95

LEARNING SYSTEM DESIGN,
Robert H. Davis, Lawrence T. Alexander and Stephen L. Yelon

Suggesting that good tenets of instruction are universal, the authors discuss principles of learning and implications for the systematic design of effective learning experiences. A wide range of examples are cited. Each chapter includes specific objectives, response frames, and a post-test to allow self-instructional feedback.

1974, 350 pages, $10.95

LEARNING THEORY, INSTRUCTIONAL THEORY, AND PSYCHOEDUCATIONAL DESIGN, Glenn E. Snelbecker

Concerned with use of psychological theory and research and development in education, this text relates learning and instructional theories to practical classroom application. It provides objective descriptions of explorations in psychology learning theory, discusses the pertinent characteristics and general importance of instructional theory and outlines ways in which sources of information can be related to improving educational practice.

1974, 544 pages, $10.95

College & University Division
McGRAW-HILL BOOK COMPANY
1221 Avenue of the Americas, New York, N.Y. 10020

Prices subject to change.

CHARLES C THOMAS • PUBLISHER

SUCCESS OR FAILURE BEGINS IN THE EARLY SCHOOL YEARS *by* **Mary Lu Kost.** *Foreword by* **Robert E. Valett.** Based on the author's premise that academic and social behavioral failures often accompany one another, she offers devices, techniques and methods for correcting the child's weaknesses. These methods, originally developed for individuals and small groups with severe learning problems, have been adapted for use in regular classroom situations. *'72, 500 pp., 44 il., cloth-$16.95, paper-$14.75*

TEACHING READING THROUGH MOTOR LEARNING *by* **Dorothy D. Sullivan** *and* **James H. Humphrey,** *both of Univ. of Maryland, College Park. Foreword by* **Robert M. Wilson.** Emphasis is placed on the use of motor-oriented activities in contrast to the more traditional ideational or verbal learning activities of reading instruction as the authors capitalize on the physical aspect of personality as an important medium of learning. The relationship between motor learning and reading and how to utilize motor activities in the reading program are included. *'73, 160 pp., 3 il., 4 tables, $7.95*

INTEGRATION OR SEGREGATION FOR THE PHYSICALLY HANDICAPPED CHILD? *By* **Susan Schmidt Dibner** *and* **Andrew S. Dibner,** *both of Duke Univ. Center, Durham, N.C.* A two-year study of handicapped children in summer camps answers the questions: Does integration provide a more "normal" life? How do adults help and hinder relationships among children? Which children should be integrated, which segregated? What are the effects of stifling the curiosity of normal children about handicaps? *'73, 228 pp., 11 tables, $8.95*

IMPROVING SCHOOL DISCIPLINE. *Edited by* **Leslie J. Chamberlin,** *Bowling Green State Univ., Bowling Green, Ohio, and* **Joseph B. Carnot,** *State Univ. of New York, Cortland. (55 Contributors)* Sociologists, students, professional writers, psychologists, teachers and educational administrators have contributed from their wide range of knowledge, experience and insights to present a realistic, workable perspective for the establishment and maintenance of school discipline. *'73, 244 pp. (7 x 10), $12.75*

ARITHMETICAL DISABILITIES IN CEREBRAL PALSIED CHILDREN: Programmed Instruction – A Remedial Approach *by* **Simon H. Haskell,** *Institute of Education, Univ. of London.* Written for teachers, educational administrators, psychologists and researchers concerned with education of children handicapped by intellectual and neurological disabilities, it is the only comprehensive account of the nature and extent of computational disorders in such children. Included are the methods for teaching arithmetic. *'73, 132 pp., 35 tables, $6.95*

301-327 East Lawrence Avenue • Springfield • Illinois • 62717

Journal of Educational Psychology

VOLUME 66 NUMBER 2

APRIL 1974

Ake
Docket 8/1 & 8/2
to the Library
[illegible] Dan
9/3/74

AMERICAN PSYCHOLOGICAL ASSOCIATION

EDITOR

JOANNA WILLIAMS, *Teachers College, Columbia University*

ASSOCIATE EDITOR

SAMUEL BALL, *Educational Testing Service*

ADVISORY EDITORS

PAUL R. AMMON, *University of California, Berkeley*
JERE BROPHY, *University of Texas at Austin*
C. VICTOR BUNDERSON, *Brigham Young University*
LEONARD S. CAHEN, *Educational Testing Service*
LOIS-ELLIN DATTA, *National Institute of Education*
FREDERICK B. DAVIS, *University of Pennsylvania*
FRANCIS J. DIVESTA, *Pennsylvania State University*
ROBERT EBEL, *Michigan State University*
DAVID ELKIND, *University of Rochester*
LAWRENCE T. FRASE, *Bell Laboratories*
N. L. GAGE, *Stanford University*
ROBERT GAGNÉ, *Florida State University*
ROBERT GLASER, *University of Pittsburgh*
MIRIAM L. GOLDBERG, *Teachers College, Columbia University*
EDMUND W. GORDON, *Teachers College, Columbia University*
VERNON C. HALL, *Syracuse University*
ARTHUR R. JENSEN, *University of California, Berkeley*
PHYLLIS A. KATZ, *City University of New York, Graduate School and University Center*
HERBERT J. KLAUSMEIER, *University of Wisconsin—Madison*
ROBERT E. LANA, *Temple University*
GERALD S. LESSER, *Harvard University*
JOEL R. LEVIN, *University of Wisconsin—Madison*
FREDERICK J. MCDONALD, *Educational Testing Service*
WILBERT J. MCKEACHIE, *University of Michigan*
FRANK B. MURRAY, *University of Delaware*
LAUREN B. RESNICK, *University of Pittsburgh*
WILLIAM D. ROHWER, JR., *University of California, Berkeley*
ROBERT ROSENTHAL, *Harvard University*
ERNST Z. ROTHKOPF, *Bell Laboratories*
S. JAY SAMUELS, *University of Minnesota*
SEYMOUR B. SARASON, *Yale University*
MELVYN I. SEMMEL, *Indiana University*
IRVING SIGEL, *Educational Testing Service*
HARRY SILBERMAN, *University of California, Los Angeles*
RICHARD E. SNOW, *Stanford University*
GEORGE G. THOMPSON, *Ohio State University*

JOURNAL STAFF

HAROLD P. VAN COTT, *Managing Editor*
ANITA DEVIVO, *Executive Editor and General Manager*
DEBORAH S. FELDER, *Supervising Technical Editor*
ANN G. CORBETT, *Technical Editor*
FORREST T. MULLINS, *Manager, Publications Business Service*
ANNE REDMAN, *Subscription Manager*

Published bimonthly in one volume per year, six issues (February, April, June, August, October, and December), by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D. C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D. C., and additional mailing offices.

The **Journal of Educational Psychology** publishes original investigations and theoretical papers dealing with problems of learning and teaching and with the psychological development, relationships, and adjustment of the individual. Preference is given to studies of the more complex types of behavior, especially in or relating to educational settings. Journal articles pertain to all levels of education and to all age groups.

Manuscripts: Submit manuscripts in duplicate to the Editor, Joanna Williams, P.O. Box 51, Teachers College, Columbia University, New York, New York 10027, according to instructions provided elsewhere in the Journal.

Copyright and Permissions: All rights reserved. Written permission must be obtained from the American Psychological Association for copying or reprinting text of more than 500 words, tables, or figures. Permissions are normally granted contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of $10 per page of APA material. Permission fees are waived for authors who wish to reproduce their own articles, or when material is reproduced in limited quantities for instructional purposes. Address requests for permission to the Permissions Office, American Psychological Association, 1200 Seventeenth Street, N. W., Washington, D.C. 20036.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1974: $24 domestic, $25 foreign, $5 single issue. APA member rate: $10. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive galley proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Department, APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., 175 Fifth Avenue, New York, New York 10010; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. Undelivered copies resulting from address changes will not be replaced; second-class mail is not forwarded unless subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

Volume 66 April 1974 Number 2

Sex Bias in the Evaluation of Professional Achievements 157
HARRIET N. MISCHEL

Project Success Environment: An Extended Application of Contingency Management in Inner-City Schools 167
HOWARD A. ROLLINS, BOYD R. McCANDLESS, MARION THOMPSON, AND WILLIAM R. BRASSELL

Concretization and Sentence Learning 179
RICHARD C. ANDERSON

Seven Aspects of Teaching Concepts 184
M. I. CHAS. E. WOODSON

Teaching Facts About Drugs: Pushing or Preventing? 189
RICHARD B. STUART

Learning from Prose Text: Effects of Readability Level, Inserted Question Difficulty, and Individual Differences 202
JACK H. HILLER

Adjunct Questions and the Comprehension of Prose by Children 212
INGRID SWENSON AND RAYMOND W. KULHAVY

Learning by Observing versus Learning by Doing 216
DOUGLAS K. CHALMERS AND MILTON E. ROSENBAUM

Effect of Type of Objective, Level of Test Questions, and the Judged Importance of Tested Materials upon Posttest Performance 225
ORPHA K. DUELL

Effects of Imagery on Learning Incidental Material in the Classroom 233
FRANK GOLDBERG

Morning-to-Afternoon Changes in Cognitive Performances and in the Electroencephalogram 238
EDMUND J. MACKENBERG, DONALD M. BROVERMAN, WILLIAM VOGEL, AND EDWARD L. KLAIBER

Effects of Feedback, Learner Control, and Cognitive Abilities on State Anxiety and Performance in a Computer-Assisted Instruction Task 247
JOE B. HANSEN

A Test of the Theory of Fluid and Crystallized Intelligence in Middle- and Low-Socioeconomic-Status Children: A Cross-Lagged Panel Analysis 255
FRANK L. SCHMIDT AND WILLIAM D. CRANO

Training Imagery Production in Young Children Through Motor Involvement. 262
WILLIAM H. VARLEY, JOEL R. LEVIN, ROGER A. SEVERSON, AND PETER WOLFF

(Continued on next page)

Diary No. 809
Date 16/9/74
Educ. Psy. Research.

Social-Emotional, Cognitive, and Demographic Determinants of Poor School Achievement: Implications for a Strategy of Intervention 267
MARTIN KOHN AND BERNICE L. ROSMAN

Determination of the Ordering Among Seven Piagetian Tasks by an Ordering-Theoretic Method 277
WILLIAM M. BART AND PETER W. AIRASIAN

Erratum 224

List of Manuscripts Accepted 254

Announcement 266

SUGGESTIONS TO CONTRIBUTORS

In preparing articles for publication in the Journal, authors are strongly advised to follow the general directions given in the *Publication Manual of the American Psychological Association* (1967 Revision). All articles must be preceded by an abstract of 100–120 words. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833). Special attention should be given to the section on the preparation of the references (p. 50). In reference lists, give journal titles in full; do not abbreviate. *All copy must be double spaced, including the references.* All manuscripts should be submitted in *duplicate*, one of which should be an original typed copy. Since the reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the identity of the author or his affiliation should be on separate pages. Every effort should be made by the author to see that the manuscript itself contains no clues as to his identity. Dittoed and mimeographed copies are not acceptable and will not be considered. Original figures must be fully prepared for reproduction according to the instructions in the *Manual*. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail and to check carefully the typing of the final copy.

Journal of Educational Psychology
1974, Vol. 66, No. 2, 157-166

SEX BIAS IN THE EVALUATION OF PROFESSIONAL ACHIEVEMENTS[1]

HARRIET N. MISCHEL

Stanford University

In a series of studies, U.S. and Israeli subjects were presented with a set of journal articles from diverse fields. The independent variables were the sex of the ascribed author (e.g., John R. Simpson, Joan R. Simpson), the male versus female association of the field in which the article was written (e.g., law versus primary education), and the sex and age of the subjects who were asked to evaluate each article. In the first study, U.S. male and female high school and college subjects showed sex bias in evaluating the journal articles by rating the same article more favorably when it was attributed to an author of one sex rather than the other sex. Judges tended to prefer authors whose sex was the same as that normative for (or strongly associated with) the professional field in which the article was written (e.g., a female author in dietetics, a male author in city planning). In a cross-cultural follow-up study, Israeli subjects did not show the evaluative biases found in the U.S. sample. Nevertheless, they had the same stereotypes as the U.S. subjects regarding the sex association of the diverse fields.

The evidence regarding the nature and extent of psychological sex differences leaves considerable room for diverse interpretations (e.g., Maccoby, 1966; Mischel & Mischel, 1971). But whether or not men and women differ substantially on many psychological dimensions and whether these differences are innate, immutable, or merely the residue of social learning, the *evaluation* of men and women seems to differ considerably (Goldberg, 1967; Gray-Shellberg, Stone, & Villareal, 1972; Horner, 1969; Pheterson, Kiesler, & Goldberg, 1971) and is reflected in highly controversial charges of rampant sexual biases that have enormous social implications (e.g., Millet, 1970). Unfortunately, the scope and nature of sex bias have been subjected more to polemics than to objective research. The present studies investigated one aspect of sex bias: the tendency to differentially evaluate achievements on the basis of the sex of the performer.

Prejudicial evaluation of women's work has been offered as an explanation for the apparent failure of women to achieve as much success as men (Klein, 1950; Scheinfeld, 1944). A study by Goldberg (1967) sought to investigate the prejudice of women against other women in the areas of intellectual and professional competence. Taking Allport's (1954) definition of prejudice as the distortion of perception and experience, Goldberg hypothesized that when confronted with an identical piece of work, women would value the professional work of men more highly than that of women. Goldberg reported a general trend on the part of his female subjects to devalue females. Women tended to rate a professional journal article more highly when it was attributed to a male author than when it was attributed to a female author, thus showing bias against their own sex.

Goldberg (1967) further hypothesized that when the professional field was one traditionally reserved for women (e.g., primary education, dietetics), the tendency of women to devalue women would be

[1] Thanks are due to Margaret Spencer and Tom Upton for help in data collection in Study 1 and to Judy Lansing for data collection in Study 3, to Antonette Zeiss for assistance with statistical analyses, and to Mordecai Nissan for facilitating access to the Jerusalem city population participating in Study 3.

Requests for reprints should be sent to Harriet N. Mischel, Department of Psychology, Stanford University, Stanford, California 94305.

lessened or even reversed. However, Goldberg reported that even in these fields, women were prejudiced against other women. Goldberg's finding of antifeminism among women is sometimes cited as if it were definitive, but a closer examination of the data he reported limits their significance. Only one out of the six fields evaluated showed a difference (i.e., preference for a male author) at less than the .05 significance level by a one-tailed test and this was for city planning, an occupation strongly associated with men for Goldberg's population. Subsequent studies also have cast some doubt on the generality of Goldberg's finding. For example, Pheterson (1969) found that sex bias was absent in a group of middle-aged uneducated women. Pheterson et al. (1971) found that female college students from the same population as Goldberg's subjects rated paintings attributed to male art-contest entrants more favorably than the identical paintings attributed to female entrants. However, when the subjects were told that the artists had been awarded prizes, they rated men and women equally.

Goldberg's (1967) data were provocative but the conclusions that may be drawn from them are limited not only by the weakness of his findings but by the fact that his subjects were restricted to female college students. What should be expected from men? Goldberg's most surprising result was that college women tended to be biased against other women even in traditionally feminine fields. Would men share this bias? Goldberg's study was executed in the mid-1960s. What changes could be expected as a result of increasing awareness, especially among young adults? Presumably there are cultural limits on the biases that Goldberg reported. More elucidation is needed on the parameters of these limits. For example, how much of an antibias factor is college education? Would the obtained biases be similar in a social structure different from our own?

The present studies were designed to investigate the foregoing questions using articles from the same fields as Goldberg had and using the same evaluative format that he had employed. Measures of sex prejudice and sex stereotypes were administered to high school and college students as well as some older adults. Male and female subjects were included in each age group so that possible sex differences in any prejudice could be noted. The measures were given to a sample of Israelis living on a kibbutz, to one of Israelis living in a large city (Jerusalem), and to U.S. subjects.

Study 1

In the first study, it was hypothesized that in traditionally masculine fields, both men and women would show bias against women. Thus it was expected that men would show the same preference for male authors in traditionally male-associated fields as did the women in Goldberg's study.

It was further hypothesized that men (unlike the women in Goldberg's study) would actually be biased in favor of women but only in certain restricted, traditionally feminine areas of work. It was reasoned that by acknowledging certain areas of proficiency for women, men can demonstrate their ostensible fairness while actually excluding or derogating women in areas reserved for males. Consider, for example, the idea that women are more sensitive in matters involving emotions and feelings, whereas men are superior in reasoning and intellectual matters, or that women are experts in keeping home and hearth, while men are best equipped to handle the affairs of the world (as President Richard M. Nixon commented to Premier Chou En-lai during his 1972 visit to Peking). Such notions may reflect a kind of "separate but equal" delegation of abilities wherein women are allowed special "female" expertise but are devalued in traditionally masculine domains.

Method

Subjects and experimenters. Subjects were 28 high school students (14 male and 14 female) and 28 college students (14 male and 14 female). The high school students were from the senior class of a large public high school drawing from a varied middle- to lower-middle-class population in Sunnyvale, California. The college students were Stanford University undergraduates. The same female Stanford undergraduate served as the main experimenter for all subjects. A male Stanford undergraduate assisted in administering the experimental procedure to the Stanford undergraduates.

Procedure. Consistent with Goldberg's (1967)

procedure, one current article was selected from the professional literature of each of four occupational fields. The fields were law, city planning, primary education, and dietetics. Goldberg's study included a preliminary rating of numerous occupational fields for their degree of association with men or women. He found that law and city planning were strongly associated with men and that primary education and dietetics were strongly associated with women. Thus, the articles in the present study were assumed to represent two extremely masculine and two extremely feminine fields.

The articles were edited and abridged to approximately 700 words each and were combined into booklets. Fictitious authors' names appeared on the first page of each article. The critical experimental manipulation involved the sex of the author as conveyed by the Christian name. Each booklet contained all four articles but, for any one article, half the booklets had a male author's name and half had a matched female author's name. Only the first name differed, e.g., John R. Simpson, Joan R. Simpson. Each booklet had two "male-authored" articles and two "female-authored" articles. Printed instructions stated:

> In this booklet you will find excerpts of four articles written by four different authors in four different professional fields. At the end of each article you will find several questions which are to be answered before you proceed to the next article. You are not presumed to be sophisticated or knowledgeable in all the fields. We are interested in the ability of high school (college) students to make critical evaluations of professional literature.

At the end of each article, the subjects answered a set of nine evaluative questions. These questions were the same for all subjects and for all articles, except of course for the author's name[2] which was included in each question. Each question, in the set of nine questions, contained five response choices, as follows:

> Based on this article, what would you judge _____'s (name of author) professional competence to be?
> 1. extremely competent, 2. above-average competence, 3. average competence, 4. below-average competence, 5. incompetent.
>
> If you were to assign a grade to _____'s (name of author) article, what would it be?
> 1. A, 2. B, 3. C, 4. D, 5. F.
>
> In her (his) appeal, was _____ (name of author)
> 1. very rational, 2. rational, 3. emotional, 4. extremely emotional, 5. alarmist.

The remaining six questions requested evaluation of the author's professional status, the value of the article for the general reader, the effectiveness of the author's writing style, and various aspects of the impact of the article on the reader (e.g., the extent to which the readers' opinions on the issues discussed were swayed by the article). For each individual, the scores for all nine rating questions were summed for each of the four fields. Thus 9 was the lowest (most favorable) possible score and 45 was the highest (most critical) possible score.

Results

As a first step, preliminary t tests were computed to compare male and female judges' ratings of the articles in the various possible author–field combinations. The results showed no sex differences approaching significance. Therefore, male and female subjects were combined in the following analyses.

The mean ratings of articles in each condition are shown in Table 1. Note that higher ratings indicate more negative evaluations. These data were submitted to a $2 \times 2 \times 2$ analysis of variance for repeated measures to test the effect of sex of fields (male and female), sex of authors, and grade level of subjects serving as judges (high school and college) on the ratings given to the articles. The results (Table 2) revealed a highly significant Sex of Field × Sex of Author interaction. Inspection of the means in Table 3 helps to clarify this interaction. The means indicated that articles attributed to male authors were preferred in the male fields (law and city planning)

[2] An interesting problem arose with regard to the inclusion of the authors' names in the evaluative questions. Goldberg's (1967) procedure was followed as closely as possible since one of the purposes of the study was to replicate and extend his findings. Goldberg reproduced only one set of questions and in that the reference was to a male author. Goldberg refers to him as, for example, Mr. Simpson. This left in doubt the title to use when the article being evaluated by the judges had been attributed to a female author. It was felt that Mrs. or Miss might have added an extra evaluative dimension that could be difficult to assess; that is, it gives additional information about the author which might influence an assessment of competence. The use of Ms. was rejected on the ground that this title is too closely associated with the women's rights movement and so might unduly sensitize the judges to the dimension of interest. The problem was resolved by using the ascribed author's first and last names only, for example, John R. Simpson and Joan R. Simpson, and omitting the Mr., Mrs., and Miss titles altogether.

TABLE 1

MEAN RATINGS OF ARTICLES IN EACH CONDITION (STUDY 1)

Subjects	Female author		Male author	
	Female field	Male field	Female field	Male field
College	19.50	23.93	22.93	21.75
High school	24.18	24.46	23.64	23.04

Note. Higher numbers indicate more negative ratings throughout the present studies.

while articles attributed to female authors were preferred in the female fields (dietetics and primary education).

The analysis of variance also showed a main effect for grade level. High school judges gave lower ratings to all the author–field combinations (Table 1), indicating that they were more critical of the articles than the college sample.

There was a significant overall interaction between all three variables: Sex of the Ascribed Author × Male versus Female Association of the Field × Grade Level of the Judges (high school and college). This three-way interaction seems to be due to the fact that the most favorable ratings were given by college judges to female authors writing in the female fields. The preference by college judges for male authors in the male fields was evident also, but not as strikingly as the preference for female authors in female fields (Table 1). Moreover, the preference patterns were shown much more clearly in the judgments of college students than of high school students, the latter showing only a general overall tendency to favor male authors. While the data in Table 1 suggest that high school judges prefer male authors to female authors regardless of the sex association of the field, further analyses (below) show that such a generalization is unjustified.

TABLE 2

ANALYSIS OF VARIANCE OF ARTICLE RATINGS AS A FUNCTION OF GRADE LEVEL, SEX OF AUTHOR, AND SEX OF FIELD (STUDY 1)

Source	*df*	*MS*	*F*
Between subjects	55		
Grade level (A)	1	182.16	4.05*
Subjects within groups (B)	54	45.00	
Within subjects			
Author's sex (C)	1	1.78	<1
A × C	1	36.17	1.02
C × B	54	35.31	
Sex of field (D)	1	30.02	<1
A × D	1	44.64	1.28
D × B	54	34.79	
C × D	1	147.88	8.77**
A × C × D	1	77.78	4.61*
C × D × B	54	16.87	

* $p < .05$.
** $p < .01$.

Separate analyses of variance were computed for each of the four fields (Table 4). Each analysis tested the effects of level of education of subject (high school or college) and sex of ascribed author on the ratings given to the articles. Significant effects were obtained only for the two female fields: dietetics and primary education. High school students preferred the primary education article when attributed to a male author. College students showed no significant differences in their preferences for this article (see Table 5).

TABLE 3

MEAN RATINGS OF ARTICLES ATTRIBUTED TO FEMALE AND MALE AUTHORS IN FEMALE AND MALE FIELDS (STUDY 1)

Author	Field	
	Female	Male
Female	21.84	24.20
Male	23.29	22.39

To further isolate the sources of the overall effects, *t* tests were performed to compare the evaluations by high school and college students on articles attributed to male versus female authors in each of the four fields. In the field of law, college students did not differentiate significantly in their evaluation of the articles, but high school students showed a marked preference for the article when it was attributed to a male author ($t = 2.48$, $p < .01$). In city planning, this pattern was reversed. College subjects markedly preferred the city planning article when it was ascribed to a male author ($t = 2.10$, $p < .05$), while the preference of high school subjects was not signifi-

TABLE 4

SUMMARY OF ANALYSES OF VARIANCE FOR EACH FIELD AS A FUNCTION OF JUDGE'S GRADE LEVEL AND SEX OF AUTHOR (STUDY 1)

Source	*df*	Law		City planning		Dietetics		Primary education	
		MS	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*
Grade level (A)	1	57.21	2.38	3.52	<1	.28	<1	481.31	16.60**
Author (B)	1	67.91	2.82	41.14	1.14	427.77	15.99**	119.62	4.12*
A × B	1	75.52	3.14	135.40	3.76	12.68	<1	98.63	3.40
Error	53	24.08		36.01		26.76		29.00	

* $p < .05$.
** $p < .01$.

cantly different for male- versus female-authored articles in this field. Thus city planning was the one field in which high school judges showed no sex bias.

In the field of dietetics, high school and college subjects both significantly preferred the article when it was attributed to a female author (college students, $t = 3.26, p < .01$; high school students, $t = 2.38, p < .05$). In primary education, the effect was as described above: high school students were biased in favor of the male author ($t = 2.78$, $p < .05$), while college students showed no significant bias.

Discussion

The obtained results indicate that while the sex of a professional does have substantial biasing effect on the evaluation of his or her achievements, the direction of these effects are variable, complex, and specific. The present study thus does not extend Goldberg's (1967) finding of professional prejudice against women to a sample of male and female judges, even when essays from identical professional fields, as well as the same evaluative format, were used. The present data, in contrast to Goldberg's conclusions, show that while an author's sex detracted from the evaluation of his or her work in some fields, it enhanced attributed status in others. Furthermore, these biases appeared to vary with the grade level of the judges, but were not affected by the judge's own sex. Female authors were preferred by both high school and college students in the field of dietetics, but high school students preferred male authors in primary education and law, while college students preferred male authors in city planning. There were no significant differences in the preferences for either a male or a female author when high school students evaluated a professional article on city planning or when college students evaluated articles in the fields of law and primary education.

Bias tended to occur in the direction of the sex appropriateness of the field. There were five instances in which one or both of the groups (high school and college) in the U.S. sample showed significant sex bias. In most of these instances, the preferred author was the one whose sex matched the sex normatively associated with that field, that is, a female author in dietetics, male authors in law and city planning. The exception to this tendency was the preference of high school students for the primary education (female field) article when it was attributed to a male author.

The sex of the judges had no effect on their evaluations of author competence in either male or female fields. This finding supported the hypothesis that there would be no differences between male and female judges evaluating professional work in male-associated fields. There was no support for the hypothesis that male judges might more strongly prefer women in fields traditionally associated with women.

TABLE 5

MEAN RATINGS OF PRIMARY EDUCATION ARTICLE (STUDY 1)

Educational level	Author	
	Female	Male
College	19.94	19.67
High school	28.43	22.87

Goldberg's (1967) finding of bias against women even in traditionally feminine fields was replicated only in the rating of the article in primary education and then only for the high school judges. Both male and female high school students preferred the primary education article when it was attributed to a male author. In the other traditionally female field, dietetics, both high school and college judges preferred the article when it was attributed to a female author. This marked preference contradicts Goldberg's result and is the only instance of sex bias on which both college and high school students agreed. In all other fields, only one group (either high school or college) showed any sex bias at all, again highlighting the specificity of the obtained prejudices.

Highly specific rather than broadly generalized sex biases were also obtained by Pheterson and her colleagues (1969; Pheterson et al., 1971). Pheterson (1969) failed to find bias in favor of males on evaluations of articles written in the fields of marriage, child discipline, and special education made by middle-aged uneducated women. In a later study, Pheterson et al. (1971) found bias in favor of males among women college students evaluating modern art that was entered in a competition but had not yet won an award. When the work had achieved an obvious success, men and women artists were rated equally. Pheterson et al. (1971) construe the results of these two studies as supporting a single interpretation. They propose that the subjects in the first study (Pheterson, 1969) perceived the very publication of an article as conferring success upon its author. Thus, they see the results of both studies as illustrating the general principle of prejudice against women until they are proved successful. These data and the data of the present study indicate far more specificity in professional evaluation than Goldberg's earlier interpretation of consistent preference for male expertise.

The differences in bias between the college and high school subjects in the present study are difficult to interpret. Nevertheless, it seems reasonable to speculate that the reduction in bias from high school to college in the field of law might reflect the subjects' increased experience with female ability in this field. College students may be coming into closer contact with the increasing numbers of women in the field of law and thus may be willing to grant females competence in this field. The increase in sex bias on the part of college students in evaluating the city planning article might be due to their clearer understanding that this occupation involves the traditionally male-associated skills of engineering and mathematics.

The high school students' bias in favor of males in the field of primary education might reflect their own dissatisfaction with the large numbers of female teachers to which they (like most U.S. children) are exposed from kindergarten through high school. College students, with this period of elementary schooling well behind them, show no sex bias in evaluating the primary education article. Perhaps their failure to prefer a female author (as they did in the other feminine field, dietetics) indicates some residual dissatisfaction.

Goldberg (1967) reported sex role stereotypes which he utilized in the design and interpretation of the results of his study. One possible explanation for the disparity between the results of the present study and those reported by Goldberg may be that judges in the present study did not share these occupational sex role stereotypes. Goldberg had based his classification on normative data obtained from a 6-point rating scale. He asked his female college subjects the degree to which they associated a list of occupational fields with men or with women. The results showed that the fields of law and city planning were most strongly associated with men, primary education and dietetics most strongly associated with women. In view of the results of Study 1, it was decided to recheck the accuracy of this classification to see whether it was applicable to current coeducational college populations.

Study 2

Method

College students (8 females and 13 males) from Stanford University were asked to indicate the

degree to which each of 10 fields was associated with men or with women. Specifically, with regard to each occupational field, the subjects were asked to rate on a 6-point scale the degree to which the field was associated either with men (at one end of the continuum), or with women (at the other end). The scale ranged from −3 (most closely associated with males) to +3 (most closely associated with females). The list included the 4 fields used in Study 1 plus 6 others: journalism, psychiatry, engineering, merchandising, art history, and linguistics.

Results and Discussion

An analysis of variance showed no overall effect of sex of subjects, but there was a strong effect for field ($F = 54.71$, $df = 9/171$, $p < .01$). The four fields used in Study 1 were rated in the expected direction: law and city planning were seen as strongly male associated, primary education and dietetics as strongly female associated.

Among the four fields included in Study 1, the mean rating for law (−2.33) indicated that it was the field most strongly associated with males. Primary education (mean rating, 1.90) was seen as the most strongly associated with females of the four fields. Nevertheless, a comparison of the rating for dietetics and city planning yielded a strong difference ($F = 129.68$, $df = 1/171$, $p < .01$), indicating that these two fields were significantly separate in their sex association in the expected direction.

The results of the second investigation clearly indicated that law and city planning are strongly male-associated fields, dietetics and primary education are strongly female-associated fields for this college sample. This finding supports the use of articles drawn from these four fields in attempting to investigate possible differential evaluation resulting from sex role stereotypes.

The correlation between sex role stereotypes and evaluative bias suggested by these studies raises the question of causation. It seems quite possible that the evaluations are rational judgments reflecting the fact that, most of the time, there is a correlation between competence and sex appropriateness. For example, since girls and women are expected to be concerned with the purchasing, preparation, and serving of food and are channeled into the professional roles of home economist and dietician,[3] it is likely that they will be more expert in the skills involved in the field of dietetics and will work better in it. Thus, much of what is referred to as "bias" in these studies may reflect reality judgments with the raters operating in the most realistic fashion possible to maximize predictive accuracy, given the situation. As societal sex roles and sex role training change, presumably so will the expectations about sex differences in professional excellence in sex-typed fields.

Study 3

The specificity found so far alerts one to the need for a more fine-grained analysis of the phenomenon of sex bias. The results suggest that the relations between sex bias and occupational sex role stereotypes are far from simple. For example, college students showed no bias in their evaluation of articles in law and primary education, although they clearly had occupational sex role stereotypes about these fields (Study 2), associating law with males and primary education with females. This suggests the possibility of stereotyping without consequent sex bias. This possibility was explored in Study 3 which examined sex bias and occupational sex role stereotypes cross-culturally. In Study 3, the procedures described above for evaluating the degree to which various occupations were associated with either sex and the differential evaluation of professional competence in selected fields were extended to an Israeli sample.

Israel, and particularly the kibbutz form of social structure, seemed a natural contrast to the U.S. sample. Both in their general philosophy and in the specific mechanisms of their social structure (especially in the kibbutz), Israelis have consciously sought to eliminate sex prejudice for many decades. It was hypothesized that the kibbutz subjects would not show bias in evaluating the articles. It was further expected that they would not show strong associations of various occupations with either

[3] These stereotypes are supported by many examples, one of which shows that males and females are shown in clearly different roles and occupations in a sample of primary reading textbooks (Jacklin & Mischel, 1973).

sex. A sample of Israelis living in Jerusalem was included so that their responses could be compared to those of kibbutz residents. It was expected that Israelis living in a city would show stronger sex association of particular occupations than Israelis living on a kibbutz but would not be sex biased in evaluating articles. It was expected that Israelis living in Jerusalem would share the sex role stereotypes of U.S. college students as a result of the similarity of the socioeconomic conditions in which they lived but that they would exhibit less sex bias as a consequence of their national ideology of sexual equality. Kibbutz residences, experiencing a social structure allowing more occupational realization of sexual equality, were expected to show less sex role stereotyping than Jerusalem residents as well as less sex bias than the U.S. groups studied.

Method

Subjects. Two groups of subjects participated in the study. One group consisted of high school youths and adults from the Kibbutz Barkai of the Artzi federation. For the evaluation of professional articles, the kibbutz sample was to have included every high school youth on Kibbutz Barkai. However, only about two thirds returned the forms. Thus, 28 subjects (14 females and 14 males between the ages of 13 and 20 years) were included in the article evaluation portion of the study. In addition, a random sample of 50 kibbutz adults (28 females and 21 males) between the ages of 25 and 48 years filled out the occupation association scale.[4]

The second group consisted of 25 students from a high school in Jerusalem. This group of 13 females and 12 males completed both the article evaluation questionnaire and the occupation association rating scale.

Procedure. The article evaluation questionnaire and the occupation association scale used in the studies conducted in the U.S. were translated into Hebrew to ensure thorough understanding on the part of the Israeli subjects. All subjects participating in the article evaluation were given the same instructions and articles (translated into Hebrew) as the U.S. subjects had received. The fictitious authors' first or "Christian" names were also translated into comparable Hebrew names (e.g., Devora Simpson, David Simpson) so that there would be no question in the judges' minds about the sex of the author. The adult kibbutz sample filled out the occupation association scale in the same way as the U.S. sample had. The occupations listed were the same as those used with the U.S. sample, with the addition of "psychologists."

The Jerusalem high school students made their ratings of the sex association of various occupations after completing the essay evaluation portion of the experiment. One student did not fill out the rating sheet correctly, therefore occupation-rating data are reported for only 24 Jerusalem students.

Results and Discussion

Sex bias. A 2 × 2 analysis of variance was computed for repeated measures to test the effect of sex of judges (male and female), sex of author (male and female), and fields (masculine and feminine) on the ratings of the articles by the kibbutz sample. The results showed no significant main effects nor interactions of the three variables. A similar statistical analysis was computed for the article evaluations of the Jerusalem sample with similar results: no significant main effects nor interactions of sex of judge, sex of authors, and fields. Thus, the ratings of the quality of the articles by both the kibbutz and the city (Jerusalem) sample showed no effect of field or author distinctions.

Occupational sex role stereotypes. An analysis of the degree to which the Israeli sample associated particular occupational fields with either males or females indicated that different fields were perceived as differently sex associated by both the kibbutz and the city samples.

For the kibbutz sample, a one-way analysis of variance with repeated measures was performed on the ratings for each field. This indicated that the different fields were perceived as differentially sex associated ($F = 142.19$, $df = 10/610$, $p < .001$). Contrasts were then computed to determine whether the four fields from which articles were evaluated were clearly sex associated for the kibbutz sample. The first contrast compared law and city planning combined with primary education and dietetics combined. It showed that these fields are indeed sex associated in the expected way ($F = 1{,}001.68$, $df = 1/610$, $p < .001$), law and city planning being strongly male associated and primary education and dietetics strongly female associated. Additional con-

[4] Although every effort was made to give the 13–20 years olds both the occupation association scale and the article evaluations, time constraints on the members of the kibbutz made it impossible.

trasts showed that primary education and dietetics are perceived as equally female fields ($F = <1$) and that law and city planning are perceived as equally male fields ($F = 1.72, ns$).

For the Jerusalem sample, a one-way analysis of variance with repeated measures was performed on the ratings for each field. This indicated that different fields are perceived as differentially sex-associated ($F = 41.93, df = 1/220, p < .001$). Next, contrasts were computed to determine whether the Jerusalem sample's ratings of law, city planning, primary education, and dietetics were sex associated in the expected way. The first contrast compared law and city planning combined against primary education and dietetics combined. It showed that the first two fields were clearly more male associated, whereas the second two fields were clearly more female associated ($F = 249.07, df = 1/220, p < .001$). An additional contrast showed that dietetics was perceived by the Jerusalem group as a more feminine field than primary education ($F = 5.04, df = 1/220, p < .05$). Primary education ranked second to dietetics in degree of female association for the 11 fields rated. Law and city planning were perceived as equally masculine ($F < 1$).

Conclusions

The results of the present three studies contradict the generalized preference for male expertise reported earlier by Goldberg (1967) and underline the subtle discriminativeness and temporal instability of sex bias. Goldberg had hypothesized that while women would prefer an article written by a male author in a male field, they would prefer a female author in a female field. His results indicated that women preferred articles attributed to male authors even in female fields and thus seemed to demonstrate a pervasive, nonspecific preference for male expertise. In contrast, in the present studies, sex bias (i.e., a differential evaluation of work on the basis of sex) appeared to be a function both of the particular field in which work was judged and of the educational level of the subject doing the evaluation. The patterns of sex bias were similar regardless of the sex of the judge. The obtained interaction between the particular field in which an author was being evaluated, the educational (or grade) level of the judge doing the evaluation, and the sex of the author indicate that sex bias was highly specific to certain fields and changed as educational level changed.

An Israeli sample (both a kibbutz and a Jerusalem group) showed no sex bias in evaluating professional journal articles. The finding that the Israeli subjects showed less sex prejudice than did the U.S. subjects might, of course, be dismissed as merely due to a failure to select occupations for which the sex role stereotypes in Israel are similar to those in the United States. Such an interpretation is untenable, however, because the results showed that the Israeli subjects had the same occupational sex role stereotypes as the U.S. subjects regarding the sex association of the four critical fields. Nevertheless, these shared stereotypes did not affect the judgments of the Israeli subjects. Unlike the U.S. subjects, they did not distinguish in their evaluations of the articles

TABLE 6

Judges' Preferences for Articles Attributed to Male (M) versus Female (F) Authors in Each Field

Judges	Female field		Male field	
	Dietetics	Primary education	City planning	Law
U.S. high school	F > M*	M > F*	M = F	M > F*
U.S. college	F > M**	M = F	M > F*	M = F
Israel (kibbutz and Jerusalem)	M = F	M = F	M = F	M = F

* $p < .05$.
** $p < .01$.

on the basis of the sex of the attributed author. A summary of the preference patterns found in both U.S. and Israeli subjects is shown in Table 6.

The shared sex occupation stereotypes found in the samples from the two cultures is congruent with several reports by Americans (e.g., Bettelheim, 1969; Spiro, 1958) and by Israelis themselves (R. R. Eiferman, personal communication, 1971) that even in Israel, men and women are thought to be more "naturally" inclined to certain different occupations. In fact, it is just this sex association of various occupations that seems to have led many of these observers to express some pessimism regarding the possibility of completely eliminating sex bias and inequality in Israel.

The present finding that sex occupation stereotypes can exist side by side with *absence* of sex bias, gives some cause for optimism. While there was no direct measure of the sex role stereotypes, if any, subscribed to by the group of kibbutz youth, it is instructive that their immediate elders (in many cases their parents) and their age counterparts in Jerusalem associate law and city planning, dietetics and primary education differentially with males and females in the same way as a U.S. college sample. There is growing evidence that beliefs, attitudes, and behavior are not necessarily consistent (e.g., Aronson, 1972; Bem, 1970). As the present data show, beliefs about occupational sex stereotypes need not produce sex bias in the judgment of competence in a particular field.

Although it is difficult to generalize from such small samples to the different cultures from which they are drawn, it seems plausible that in a culture in which women do have more professional opportunities and are seen as more equal to men in their abilities, there would be little evaluative sex bias. In Israel, the availability of, and experience with, competent women in a greater variety of fields (e.g., in the army and as head of the government) may break down biases in practice even though sex role stereotypes, remarkably similar to those found in the United States, continue to exist at the level of attitudes. Finally, it may be speculated that in studies such as the present three and those of Goldberg (1967), Pheterson (1969), and Pheterson et al. (1971), the phenomenon called sex "bias" may partly reflect reality judgments about the differential probability of success for males and females within particular fields. These sex differences in the objective probability for success within particular fields may be mirrored by the stereotypes and "biases" shared by members of the culture.

REFERENCES

Allport, G. W. *The nature of prejudice.* Cambridge, Mass.: Addison-Wesley, 1954.

Aronson, E. *The social animal.* San Francisco: Freeman, 1972.

Bem, D. J. *Beliefs, attitudes and human affairs.* Belmont, Calif.: Brooks/Cole, 1970.

Bettelheim, B. *The children of the dream.* New York: Macmillan, 1969.

Goldberg, P. A. Misogyny and the college girl. Paper presented at the meeting of the Eastern Psychological Association, Boston, April 1967.

Gray-Shellberg, L., Stone, S., & Villareal, S. A comparison across time of the double standard in career conflict situations. Paper presented at the U.C.L.A. Symposium on Sex Roles and Sex Differences, Los Angeles, May 1972.

Horner, M. Fail: Bright women. *Psychology Today,* 1969, **3,** 36–38, 62.

Jacklin, C. N., & Mischel, H. N. As the twig is bent—Sex role stereotyping in early readers. Unpublished manuscript, Stanford University, 1973.

Klein, S. V. The stereotype of femininity. *Journal of Social Issues,* 1950, **6,** 3–12.

Maccoby, E. E. Sex differences in intellectual functioning. In E. E. Maccoby (Ed.), *The development of sex differences.* Stanford: Stanford University Press, 1966.

Millet, K. *Sexual politics.* New York: Doubleday, 1970.

Mischel, W., & Mischel, H. The nature and development of psychological sex differences. In G. S. Lesser (Ed.), *Psychology and educational practice.* Glenview, Ill.: Scott, Foresman, 1971.

Pheterson, G. I., Female prejudice against men. Unpublished manuscript, Connecticut College, 1969.

Pheterson, G. I., Kiesler, S. B., & Goldberg, P. A. Evaluation of the performance of women as a function of their sex, achievement, and personal history. *Journal of Personality and Social Psychology,* 1971, **19,** 114–118.

Sheinfeld, A. *Women and men.* New York: Harcourt, Brace, 1944.

Spiro, M. E. *Children of the kibbutz.* Cambridge, Mass.: Harvard University Press, 1958.

(March 26, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 167-178

PROJECT SUCCESS ENVIRONMENT: AN EXTENDED APPLICATION OF CONTINGENCY MANAGEMENT IN INNER-CITY SCHOOLS[1]

HOWARD A. ROLLINS[2] AND BOYD R. McCANDLESS

Emory University

MARION THOMPSON AND WILLIAM R. BRASSELL[3]

Atlanta Public Schools, Georgia

Sixteen black and white inner-city public school teachers who were trained to use positive behavior contingencies did so for one academic year with a total of 730 Afro-American disadvantaged pupils from the first through the eighth grades. Compared with matched control teachers and classes, these 16 teachers showed higher incidences of positive reinforcement and lower incidences of punishment. The experimental classes were less disruptive and more on task. They gained more in both IQ and school achievement. Inner-city teachers can be trained to employ positive techniques of behavior management; they like and use such training; and public school pupils profit dramatically from restructuring their learning environment.

Both black and white children from low socioeconomic backgrounds are failing to gain an adequate education in the nation's central-city schools (e.g., Coleman et al., 1966; Dittman, 1967; Kvaraceus, 1965, McCandless, 1967, 1970). In fact, the educational achievements of these children have been repeatedly documented as dismal. As a group, they fall further and further behind their economically advantaged, suburban peers with each year of schooling (e.g., McCandless, 1970).

The list of variables offered as explanations for the academic plight of the inner-city child appears endless (c.f., Bronfenbrenner, 1974). However, as Becker, Engelmann, and Thomas (1971), among others, suggest, etiology may be less important than the academic environment in which these children are placed. Our schools are designed to build successively year after year upon skills acquired by the children in previous years. If at any point a child has not acquired the appropriate prerequisite skills, failure is likely. For inner-city children, such failures often occur early, since they typically enter school poorly prepared to handle both the standard public school curriculum and the middle-class format of the classroom. Further, a history of failure may promote expectations of failure which in turn make actual failure more likely. Thus, inner-city children are forever behind, confused, and as a consequence, probably lose all interest in undertaking new academic tasks. As a result, inner-city classrooms are filled with unhappy, restless children who are relatively uninvolved in academic work and often are highly disruptive.

[1] This study was conducted under the auspices of the Georgia Department of Education, which provided funds allocated under Title III, Public Law 89-10. Readers will note that incredible amounts of effort and goodwill on the parts of many were necessary for this study to be conducted. The Project staff, the teachers involved in the classrooms, and the principals of the different schools deserve at least as much credit as the authors. Thanks should also be extended to all the children and youth who took part.

[2] Requests for reprints should be sent to Howard A. Rollins, Psychology Department, Emory University, Atlanta, Georgia 30322, or to Marion Thompson, Director, Project Success Environment, 210 Pryor Street, Atlanta, Georgia 30303.

[3] Now at Western Carolina Center, Morgantown, North Carolina.

If this analysis is correct, then one logical course of action is to replace failure with success. To guarantee inner-city children success, the school curriculum must be modified to provide opportunities for success by matching the material presented to the level at which these children function, and teachers must be trained to emphasize success while minimizing failure.

Contingency management is one vehicle available for providing inner-city students with success experiences. Based on the principles of operant conditioning, contingency management has proved to be an effective motivational system (e.g., Staats, 1964; Staats & Staats, 1963). Both social and token reinforcements have been used in classroom settings to accelerate appropriate behaviors, such as attending behavior, and to decrease disruptive behaviors. Praise and teacher attention have provided adequate incentives for many students to perform effectively (e.g., Becker, Madsen, Arnold, & Thomas, 1967; Harris, Wolf, & Baer, 1964; Madsen, Becker, & Thomas, 1968; Zimmerman & Zimmerman, 1962). Token reinforcements—tangible objects or symbols which, when exchanged for a variety of other objects such as edibles or playthings, acquire reinforcing power themselves—have also proved effective in modifying pupil behavior (e.g., Kupers, Becker, & O'Leary, 1968; McLaughlin & Malaby, 1971; O'Leary, Becker, Evans, & Saudargas, 1969).

Other investigators have shown that academic achievement can be accelerated through manipulating contingencies. In a series of studies, Staats (1964; Staats, Minke, Finley, Wolf, & Brooks, 1964; Staats & Staats, 1963) reported significant gains in reading achievement using token reinforcement. Staats and his colleagues (e.g., Staats & Staats, 1963), among other relevant theoretical projections, suggest "work" rooms in schools where token reinforcers are earned for learning activities and other rooms and situations where the child receives the reinforcement that backs up the tokens. This was a key strategy employed in the present study. Among others, Clark and Walberg (1968) and Wolf, Giles, and Hall (1968) have reported similar results. Teachers have also been trained to manipulate contingencies in the classroom, and the results have been encouraging (e.g., Hall, Lund, & Jackson, 1968; Hamblin, Buckholdt, Ferritor, Kozloff, & Blackwell, 1971; Thomas, Becker, & Armstrong, 1968).

The following points should be made about the present study. First, it has been clearly demonstrated, and is supported by our own observations made early in the course of conducting this study, that teachers of inner-city pupils typically employ negative and even punitive methods as their major incentive technique for behavior control and academic learning. Second, it is clear that behavior modification is not necessarily a positive technique, but can be and often is accomplished by means of aversive incentives. Third, most of those reporting in the literature, either by themselves or by way of specialists trained by them, have been interested in finding whether the behavior modification technique worked but have been less interested in training classroom teachers in its use.

In the present study, we have worked to move teachers from the employment of a preponderance of negative to a preponderance of positive incentives. Appropriate behavior is rewarded, inappropriate behavior is ignored, and almost no aversive incentives are used. Also, our emphasis has been on the pre-service and in-service training of teachers rather than on that of specialists in the use of positive behavior modification. Further, most behavior modification investigators who have reported in the literature have worked with individual students or small groups for limited periods of time, such as six weeks. In the present study, a contingency management technique was implemented in a large number of inner-city classrooms from the first to the eighth grades for an entire academic year.

Entitled "Project Success Environment," the pilot 1970–1971 study included eight experimental classes with appropriate comparison classes at the first-, second-, third-, and seventh-grade levels (c.f. Thompson, Brassell, Persons, Tucker, McCandless, & Rollins, 1973). Following this initial devel-

opmental effort and its encouraging results, the program was expanded to include twice the number of students within a wider age range during the second year of operation, 1971–1972. A reasonably rigorous experimental design was also incorporated in the second year.

As a group study, Project Success Environment was not designed as an exercise in scientific analyses of behavior. It purpose, rather, was to answer an actuarial question of the sort suggested by Baer (1971): Can behavior modification solve the referring social problem, which has been analyzed into two sets of behavior—those too high and those too low in rate? The central question in this study is whether teachers can be trained to use the techniques made available through behavioral analysis to provide large numbers of students from economically disadvantaged backgrounds with some modicum of individual success.

Method

Subjects and Setting

Three hundred and sixty-seven male and 363 female black pupils enrolled in a public middle school and three of its feeder elementary schools participated in the study. The four schools are located on the fringe of inner-city Atlanta and are characterized by substandard educational achievement and a high proportion of pupils from low-income families; for example, 54.8% of the pupils in one of the elementary schools are from families with annual incomes of $2,000 or less.

The subject population during the second year of the study (the present data) consisted of pupils in the first, second, third, fourth, sixth, and eighth grades. The total sample was divided into an experimental group and a control group of 16 and 14 classes, respectively, with from 22 to 25 pupils per experimental class and from 25 to 28 pupils per control class. The control classes at the elementary level were in a nearby elementary school, while those at the middle-school level (sixth and eighth grades) were in the same school. Of the 355 experimental subjects, 154 were exposed to the treatment over a period of two consecutive years. The rest were involved in the study for the second year alone, so that the percentage of "two-year" pupils in the experimental classes during the second year ranged from 0% in seven classes to 81% in two classes. Thirteen experimental and 10 control teachers were black, the others were white.

Although control subjects were identified during both years of the study, those for whom data are reported here were selected just before the beginning of the second academic year. With the exception of the first-grade pupils, the experimental and control subjects were matched on the basis of reading scores obtained the previous April on the Metropolitan Achievement Tests.

Procedures

Behavior modifiers. The teachers of the 16 experimental classes served on a voluntary basis as behavior modifiers within the framework of the public school setting. Eight of the teachers participated in the study from its inception, while the remaining 8 participated only during the second year. The 14 control teachers were selected at the beginning of the second year by their respective principals from the available faculty at the appropriate grade levels. Most of the experimental and control teachers were female, with previous classroom experience of from 1 to 13 years. A paraprofessional aide was available to each experimental and control teacher for approximately 90 minutes per day to assist with clerical and logistical tasks.

Teacher training. Most of the training for the experimental teachers was accomplished in a three-week workshop during the summer preceding the second academic year. (The eight experienced teachers had also participated in a similar three-week seminar in the first year.) The workshop was conducted by three psychologists and three educators and was designed to provide instruction in the theory and practical application of operant conditioning and to involve the teachers in planning for the classroom implementation of behavioral management procedures and various curricular activities.

During the mornings, the teachers participated in discussion sessions that were primarily focused on readings in behavior modification from *Teaching: A Course in Applied Psychology* (Becker, et al., 1971) and other sources. The teachers then had the opportunity to apply behavioral management principles in classroom settings while being observed by their peers, and a videotape recording was made of them, which was to serve as the basis for further discussion in classroom management. The teachers were also exposed to systematic classroom observation by collecting data in actual classrooms using the procedures and forms that trained observers would use later in their classrooms. In addition, each teacher shared in the identification of the pupil behaviors to be modified during the following year and in the establishment of a token economy to support the behavior modification effort.

The afternoons during the workshop were devoted to curriculum planning, especially for the initial weeks of school. Emphasis was placed on formulating behavioral objectives for pupils, developing and using individualized instruction techniques, using programmed reading materials and academic diagnostic instruments, and establishing and maintaining a specific classroom arrangement.

Throughout the year, an experienced behavioral management technician was available at least twice weekly to assist each teacher with current problems in classroom management. Individual in-service sessions concerning curriculum implementation were also conducted weekly by two curriculum coordinators, one of whom concentrated on the elementary curriculum and the other on the middle-school curriculum.

Target behaviors. For the first six to eight weeks of school, the emphasis was on the positive reinforcement of desirable classroom conduct in an effort to increase the frequency of on-task behavior and to decrease the frequency of disruptive behavior. On-task behavior was defined as apparent attention to assigned academic tasks, while disruption subsumed any unsolicited behavior serving to distract pupils from academic tasks, for example, physical contact or inappropriate social conversation among pupils. Appropriate classroom behavior in the elementary classes was stipulated by the following set of conduct rules agreed upon by the elementary teachers during the summer workshop: (*a*) stay in your seat, (*b*) work hard, (*c*) pay attention, and (*d*) raise your hand to speak. The rules agreed upon by the middle-school teachers were as follows. (*a*) pay attention, (*b*) have necessary tools for work, (*c*) stay on task, and (*d*) raise your hand for recognition.

Although these behavioral guidelines were common across classes, each teacher was encouraged to interpret them according to her individual teaching style and to relate her precise interpretation of the guidelines to her pupils on the first day of school. Some teachers chose to specify that their pupils remain in their seats except when granted permission to move; others indicated to their pupils that they could circulate freely within the classroom, provided they were engaged in an academic task. In any event, the teachers were consistent within their respective classrooms in the specification and execution of behavioral contingencies. Once desired conduct was established within a class, the emphasis shifted partially to the reinforcement of academic behavior. For most classes, this shift to the reinforcement of academic behavior occurred no later than the third week of school.

Behavior management. The teachers adhered to a basic premise of "ignore and praise;" that is, they attempted to attend and reward suitable behavior while disregarding improper conduct. Behavioral management, then, was implemented overtly by means of positive reinforcement, although the possibility exists that the teachers may have also used subtle forms of punishment, such as the withdrawal of attention. In addition to praise and other forms of social reinforcement, the teachers relied heavily on a token system in which checkmarks on "reward cards" and tickets were dispensed in the elementary and middle-school classes, respectively. Since the elementary classes were self-contained, the elementary pupils were exposed to the behavioral contingencies throughout each school day. The middle-school classes, however, were taught by teams (three teachers per team in the sixth grade and four in the eighth grade), so that the pupils were exposed to the contingencies for approximately four hours daily during the mornings while they attended the basic classes (reading, mathematics, social studies and, in the eighth grade, science) taught by the experimental teachers. During the afternoons, the middle-school pupils attended nonexperimental exploratory classes, such as music, art, and home economics.

Throughout the first day of school and for several days thereafter, immediate primary reinforcement (M & Ms and hard candy) was paired with praise and token reinforcement contingent upon approximations of desired social conduct, including such behaviors as simply coming to school and sitting at a desk. Enough tokens were distributed within the first two days for every pupil to exchange them for a variety of backup reinforcers, including both inexpensive "fun" items and school supplies. During the initial two weeks of school, reinforcement was dispensed on a generally continuous and predictable basis, but as the desired behaviors were gradually shaped, the tokens were dispensed on more intermittent, less predictable schedules. In order to generate more intermittent schedules and to provide a mechanism for delay of reinforcement, the number of checkmarks required to complete a reward card for the elementary pupils increased progressively from 25 to 150. In the middle-school classes, not only were less expensive items (in terms of tokens) replaced by more expensive items, but the prices of various items were increased so that, for instance, an item which could be obtained for 5 tickets on the first day of school was worth 25 tickets by the end of the second week.

During the third and fourth months, all tangible rewards, with the exception of certain school supplies, were replaced by activity reinforcers in that the pupils traded tokens for access during school hours to the activity room supervised by the paraprofessional aides. The activity rooms at the elementary schools were stocked with such items as games, toy cars, comic books, sewing kits, dolls, Tinker Toys, and Lincoln Logs, while the middle-school activity room contained not only games, magazines, and puzzles, but also records for listening and dancing, as suggested by a student committee. The elementary pupils were also able to exchange tokens for the privilege of assisting the teacher in such capacities as playground monitor, chalkboard monitor, and "mini-teacher" or tutor.

Although there had been some concern that the pupils would find the transition from tangible to intangible rewards unpalatable, there were no apparent detrimental effects, perhaps because there was a four- to six-week overlap between the two types of reinforcement and because the pupils

were told of the impending change six weeks in advance. Immediately prior to the final conversion to activity reinforcers, auctions were held in each classroom to dispose of the remaining tangible rewards. From this point through the remainder of the year, tokens could be exchanged only for activity reinforcers and a limited assortment of school supplies, such as pencils and notebooks.

Classroom arrangement. A classroom arrangement, consisting of a mastery center for instruction and five academically oriented interest stations, served to structure the instructional program and concomitantly to free the teachers for more interaction with individual pupils and small groups. Within the mastery center, the pupils were divided into three ability groups in which they received instruction and completed academic assignments. While one group received instruction and the second completed assigned tasks, the third group visited the various interest stations that were designed to foster individual and small group exploratory behavior without direct teacher intervention. The five stations included a library station with books, magazines, and newspapers; an art station with a variety of paints, crayons, and other materials; a communications station with a language master, phonograph, and tape recorder; and an exploratory station with an assortment of science materials keyed to the instructional program; and a games and puzzles station equipped primarily with academically related materials. The materials at the stations were changed or rotated among the classrooms at least weekly by the paraprofessional aides.

Control classes. All control classes were conducted in a traditional manner with a single teacher managing each class in a lecture format. Control teachers had access to numerous academic materials but seldom used materials other than those prescribed by the school system. None of the control teachers received any formal training in the use of contingency management procedures.

Behavioral Observations

Five paraprofessional data gatherers systematically observed teacher and pupil behavior in each class for 45 minutes daily during the first three weeks of school and twice weekly during the remainder of the year. Each observer gathered data in both the experimental and the control classes during observation sessions, which varied from morning to afternoon and occurred only during periods of academic activity.

During each 45-minute observation session, the relevant behaviors were observed three times in structured 15-minute sequences in order to obtain more typical behavioral samples. Within a single 15-minute sequence, data were obtained using three different procedures to observe and record teacher reinforcement and punishment, pupil disruption, and pupil attention, in that order.

Positive reinforcement and punishment. The teacher alone was observed for 5 minutes in each sequence (for a total of 15 minutes during the entire 45-minute observational session). The data gatherer counted and classified every instance of teacher-administered positive reinforcement according to the nature of the reinforcer (tangible or nontangible) and according to the nature of the behavior being reinforced (academic or conduct). The average number of positive reinforcements administered per student in a 15-minute period constituted a criterion measure, which was obtained by dividing the total number of reinforcements administered by the number of pupils present during the observation session. A second criterion measure consisted of the total number of instances of punishment.

Teacher behaviors recorded as positive reinforcement included verbal praise, positive physical contact, the granting of privileges, and the administration of tangible rewards such as candy or tokens (which were administered only in the experimental classes). Punishment included criticism implied explicitly or implicitly through threats of consequences, voice tone, or facial expression; aversive physical contact with pupils; withdrawal of pupil privileges; and isolation of pupils.

Disruption. During the second 5-minute period in each 15-minute sequence, the data gatherer ignored the teacher while continuously scanning the entire class for instances of disruptive pupil behavior. In general, disruption encompassed any unsolicited pupil behavior serving to distract other pupils from academic tasks, such as talking or being out of one's seat without permission; generating loud noises; or disturbing other pupils either verbally, by means of physical contact, or by handling another pupil's possessions. A single pupil could not be observed for disruption more often than once every 10 seconds. The criterion measure was the average number of disruptions per pupil per 15 minutes, obtained by dividing the total number of disruptions recorded by the number of pupils present during the observation session.

Attention. Attentive behavior was observed during the final 5-minute period of the 15-minute sequence. One third of the pupils assigned an academic task were observed during each of the three 5-minute periods, with the focus being on attention; therefore, each pupil was observed for attentive behavior one time only for 20 seconds within the entire 45-minute observation session. The data gatherer recorded the number of seconds during which the pupil was off task; that is, during each 20-second interval, the behavior of one pupil was observed and the amount of time apparently devoted to other than academic tasks was recorded. Each pupil observed was classified as involved (0–5 seconds off task), medium involved (6–15 seconds off task), or uninvolved (16–20 seconds off task). The criterion measure was the percentage of time on task for the entire class, calculated by adding the number of pupils classified as involved to one half of the number classified as medium involved, then dividing the sum

by the total number of pupils observed, and multiplying the quotient by 100.

Interrater reliability. Reliability coefficients were obtained periodically for the five data gatherers by comparing their observations with the simultaneous observations of one of the behavioral management technicians. Most of the resulting coefficients were above .80. The median coefficients (over 12 reliability checks) for reinforcement, punishment, disruption, and attention were .94, .78, .90, and .88, respectively.

Academic aptitude and achievement. One of the primary goals of the program was to accelerate both academic aptitude and achievement. In order to assess these objectives, the short forms of the California Test of Mental Maturity and the California Achievement Test were administered to all experimental and control subjects in mid-September and again in early May. The latter test was administered at reading level rather than grade level, since testing during the first pilot year of the study had indicated that these children frequently performed at chance when tested at grade level.

Results

Behavioral Observation

The teacher training program, among other things, was designed to increase the frequency of teacher-administered reinforcements and to reduce the frequency of teacher-administered punishments. The effects of the training program on teacher reinforcements are given in Figure 1 relative to controls across the entire school year. Each data point in Figure 1 is an average of the observations within a one- or two-week period. It is evident in Figure 1 that experimental teachers administered more reinforcements per student than control teachers. In fact, the experimental teachers essentially doubled the rate of reinforcement administered by control teachers. The number of reinforcements delivered by project teachers dropped over the course of the school year, reflecting progress toward an intermittent schedule. However, even at the end of the year, project teachers reinforced their students at more than twice the rate of control teachers.

A 2 × 2 × 18 (Frequency of Reinforcement Delivery with Treatment Group × Elementary versus Middle-School Grade Level × Observation Interval) analysis of variance was performed. Consistent with the above observations, experimental teachers administered reliably more reinforce-

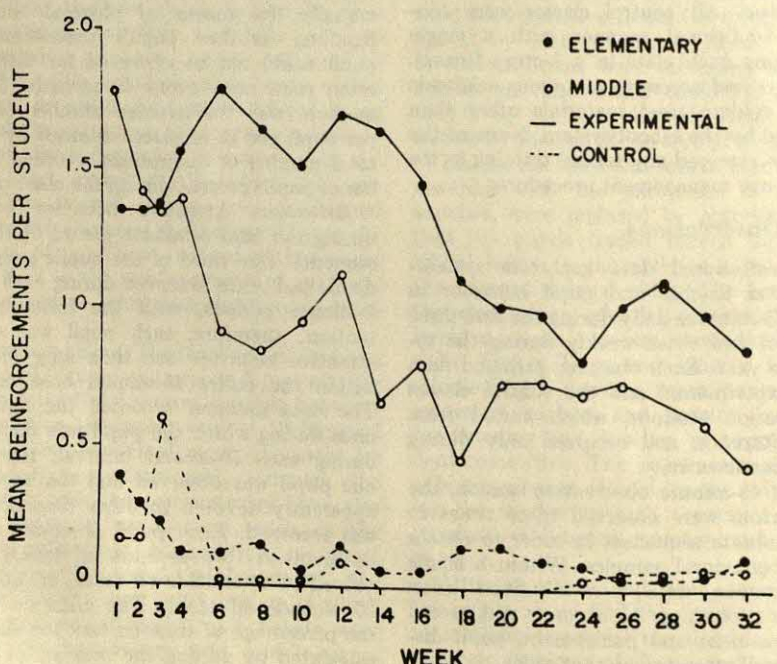


Figure 1. Mean number of reinforcements per student over weeks as a function of treatment condition and grade level.

ments than control teachers ($F = 36.98, df = 1/26, p < .01$). In addition, the Treatment Group × Observation Interval interaction was reliable ($F = 4.63, df = 17/442, p < .01$), reflecting the reduced separation of experimental and control teachers across the school year. There were no reliable effects of grade level.

Figure 2 is a representation of the average number of punishments per 15-minute interval delivered by project and control teachers by grade level. As shown in Figure 2, project teachers were much less punitive than control teachers, particularly during the first few weeks of school. For the first two weeks, project teachers at both elementary and middle schools administered fewer than one fourth as many punishments as control teachers. The rate of punishment declined over the school year for all groups. At the elementary level, project and control teachers punished at about the same level during the last two weeks, while controls at the middle-school level continued to administer more punishment throughout most of the year. In line with these observations, an analysis of variance yielded significant effects for treatment ($F = 3.35, df = 17/442, p < .01$). In addition, grade level and weeks interacted ($F = 4.20, df = 17/442, p < .01$), indicating the greater decline in punishment rate at the elementary level than at the middle-school level.

These data demonstrate quite clearly that the project teachers' behaviors were appropriately modified by the summer and in-service training. As a result of these changes, students in project classes received a rather massive dose of contingent "success" and minimal punishment. The alterations of classroom environment produced some marked changes in pupil behavior as evidenced by the data for disruptions and task involvement presented in Figures 3 and 4. As indicated in Figure 3, students in project classes emitted fewer disruptions than chil-

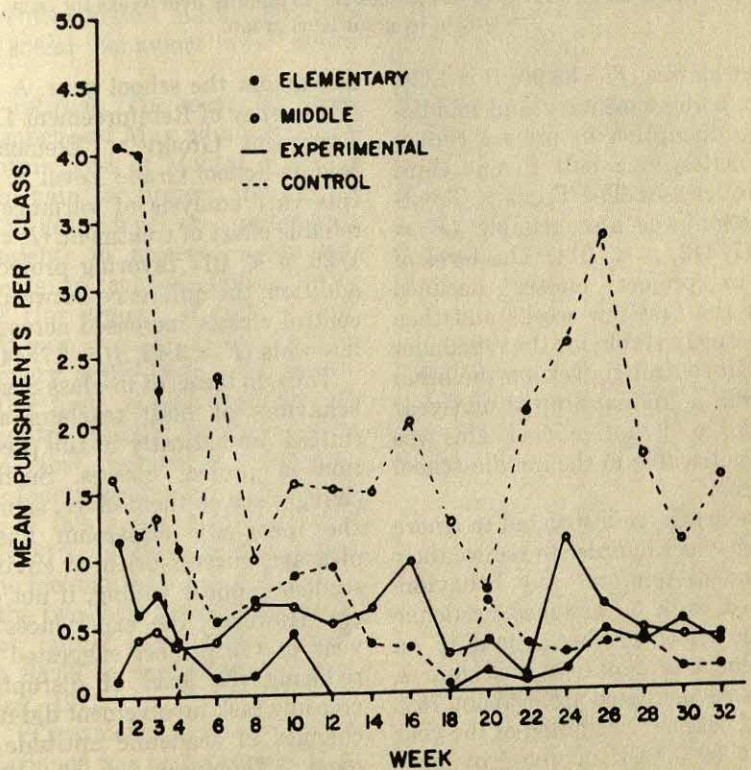


FIGURE 2. Mean number of punishments delivered per 15 minutes as a function of treatment condition and grade level.

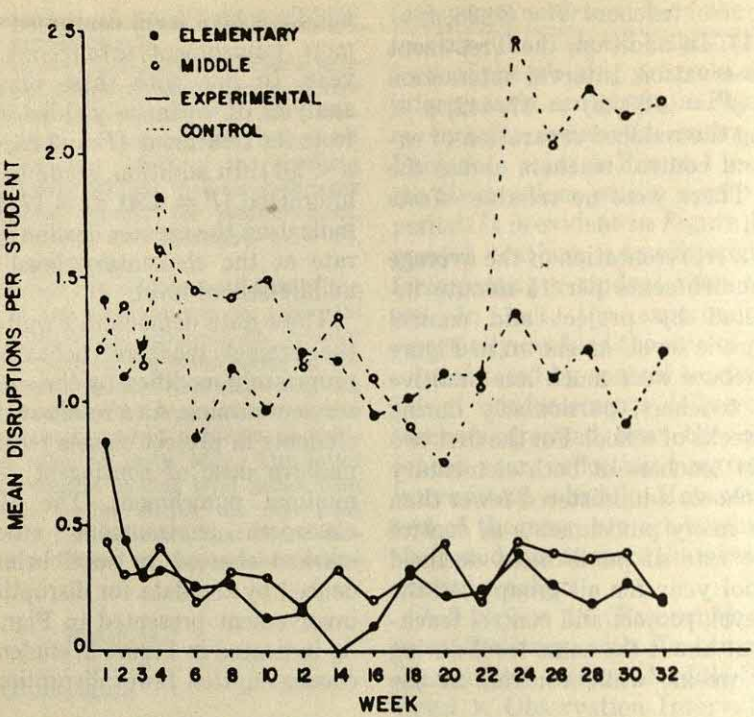


FIGURE 3. Mean number of disruptions per 15 minutes over weeks for each treatment by grade level group.

dren in control classes ($F = 33.99, df = 1/26, p < .01$). At both elementary and middle-school levels, disruption in project classes was approximately one half to one third that of control classes. The Trials × Treatment interaction was also reliable ($F = 2.30, df = 17/442, p < .01$). The level of disruption in project classes declined slightly over the first few weeks and then remained relatively stable for the remainder of the year. In control classes, on the other hand, there was a gradual drop at mid-year and then a rise in disruptive level. This was particularly noticeable in the middle-school control classes.

Project teachers were instructed to ignore disruptive behaviors in order to reduce their frequency and to reinforce any behaviors that evidenced work on assigned academic material. The effect of this procedure on task involvement is apparent in Figure 4. Project students increased from about 75% task involved near the beginning of the year to more than 90% task involved by mid-October. Control students, on the other hand, remained from 65% to 75% involved throughout the school year. A 2 × 2 × 18 (Frequency of Reinforcement Delivery with Treatment Group × Elementary versus Middle-School Grade Level × Observation Interval) analysis of variance indicated a reliable effect of treatment ($F = 60.89, df = 1/26, p < .01$) favoring project classes. In addition, the difference between project and control classes increased across observation intervals ($F = 3.42, df = 17/442, p < .01$).

Thus, in terms of in-class observation, the behaviors of both teachers and students shifted dramatically in the predicted direction in project classes. Such behavioral changes are, in themselves, significant, since the inner-city classroom has become a pleasant, success-oriented environment and students appear willing, if not eager, learners. However, the experiences of the first year of the project suggested that simply reducing the level of disruption and increasing task involvement did not guarantee changes in academic aptitude or achievement (Thompson, et al., 1972). Consequently, in the second year, teachers were encouraged to reinforce evidence of aca-

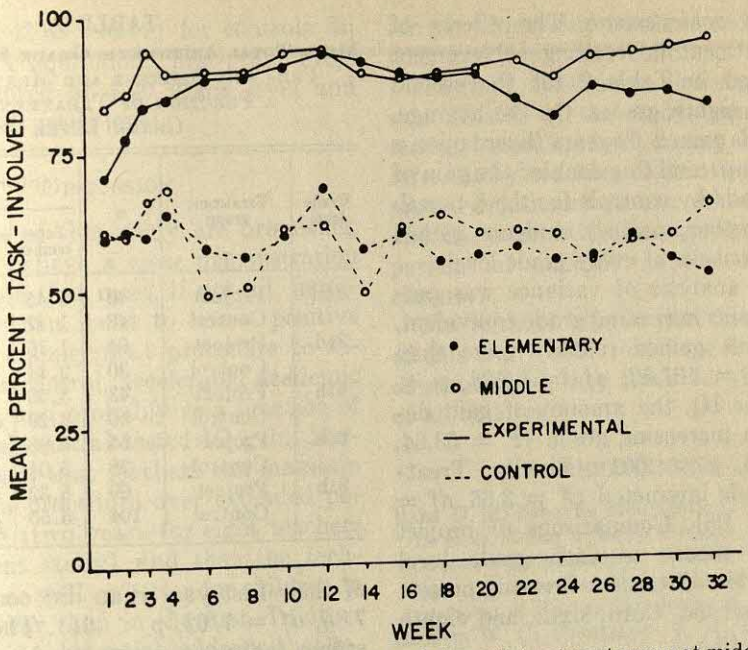


FIGURE 4. Mean percentage of pupils on task for each treatment group at middle and elementary grade levels.

demic achievement almost exclusively, once appropriate social behaviors were established.

Academic aptitude. The mean IQ scores for the September and May administrations of the California Test of Mental Maturity and the mean gain for project and control classes at each grade level are presented in Table 1. Students in project classes outgained students in control classes by a factor over 2 (for project, M gain = 5.98; for control, M gain = 2.51). In fact, project students at every grade level, with the exception of the first, achieved greater IQ gains than control pupils. As indicated in Table 1, the most impressive change occurred at the fourth grade. The outstanding performance of this group may be due to the fact that 81% of these children were exposed to the technique for two consecutive years. Over the two-year period, the fourth-grade project pupils have gained 20 IQ points: from 85.69 in September, 1970, to 105.56 in May, 1972.

A 2 × 6 (Treatment × Grade Level) analysis of variance was performed on gains in IQ. Project pupils gained reliably more than controls ($F = 12.14$, $df = 1/602$, $p < .01$). There was some drop in the amount of gain from the first to the eighth grade ($F = 3.85$, $df = 5/602$, $p < .01$). Finally, Grade × Treatment interacted ($F = 5.90$, $df = 5/602$, $p < .01$), indicating the large reversal in gain at the first-grade level. Project students reliably outgained controls by specific comparison for third, fourth, sixth, and eighth grades.

TABLE 1

MEAN TOTAL IQ (CALIFORNIA TEST OF MENTAL MATURITY) AT PRE- AND POSTTEST AND MEAN GAIN AS A FUNCTION OF TREATMENT AND GRADE LEVEL (September to May)

Grade	Group	n	Pretest	Posttest	Gain
1	Project	35	98.60	98.49	−.11
	Control	34	86.60	96.18	9.58
2	Project	38	86.03	94.21	8.18
	Control	44	86.11	92.00	5.89
3	Project	61	88.28	95.54	7.26
	Control	35	90.54	92.94	2.40
4	Project	36	91.69	105.56	13.87
	Control	43	86.70	88.17	1.47
6	Project	57	85.70	91.40	5.70
	Control	49	85.76	86.27	.51
8	Project	76	71.63	74.67	3.04
	Control	106	73.86	74.08	.22

Academic achievement. The effects of project treatment on reading achievement are presented in Table 2 for the second through the eighth grades. On the average, project pupils gained .69 years (based upon a 10-month year) and this doubled the gain of .34 years made by controls for the 8-month interval. Further, project students gained more than controls at every grade level.

A 2 × 5 analysis of variance was performed on gains in reading grade equivalent. Project pupils gained reliably more than controls ($F = 167.60$, $df = 1/528$, $p < .001$). As for IQ, the amount of gain decreased with increasing grade ($F = 53.64$, $df = 4/528$, $p < .001$). Finally, Treatment × Grade interacted ($F = 3.65$, $df = 4/528$, $p < .05$). Comparisons of project and control means at each grade level yielded reliable differences favoring project students at second, third, sixth, and eighth grades ($p < .01$).

Children in first-grade classes were given the achievement tests only at the end of the year. However, since IQ scores were obtained in September and since project students scored higher on this pretest, an analysis of covariance was performed comparing year-end reading scores for first-grade students with IQ as the covariate. This analysis indicated that project first graders had better reading skills at the end of their first year than did controls ($F = 7.85$, $df = 1/62$, $p < .01$). The mean raw scores (adjusted scores are in parentheses) in reading were 60.94 (59.22) for project pupils and 46.70 (48.71) for controls. These mean raw scores translate to 1.80 and 1.55 in grade equivalent for project and control groups, respectively.

TABLE 2

MEAN TOTAL READING GRADE EQUIVALENT ON CALIFORNIA ACHIEVEMENT TEST AT PRE- AND POSTTEST AND MEAN GAIN AS A FUNCTION OF TREATMENT AND GRADE LEVEL

Grade level	Treatment group	*n*	Reading		Gain
			September	May	
2nd	Project	46	1.54	2.72	1.18
	Control	40	1.46	2.11	.65
3rd	Project	61	1.87	2.48	.61
	Control	34	2.19	2.50	.31
4th	Project	46	3.28	3.90	.62
	Control	23	2.74	3.31	.57
6th	Project	57	4.55	5.05	.50
	Control	37	4.51	4.90	.39
8th	Project	77	4.67	5.29	.62
	Control	117	4.90	5.08	.18

TABLE 3

MEAN TOTAL ARITHMETIC GRADE EQUIVALENT AT PRE- AND POSTTEST AND MEAN GAIN AS A FUNCTION OF TREATMENT AND GRADE LEVEL

Grade level	Treatment group	*n*	Arithmetic		Gain
			September	May	
2nd	Project	46	1.45	2.01	0.56
	Control	43	1.47	1.87	0.40
3rd	Project	64	1.76	2.34	0.58
	Control	30	2.14	2.56	0.42
4th	Project	43	3.58	4.26	0.68
	Control	50	3.29	3.72	0.43
6th	Project	55	4.98	5.94	0.96
	Control	28	5.04	5.34	0.30
8th	Project	69	5.32	5.83	0.51
	Control	104	5.55	5.95	0.40

Results for arithmetic achievement for Grades 2 through 8 are presented in Table 3. Overall, project pupils gained .65 years in mathematics (based upon a 10-month school year) as compared to a .39 gain for control pupils. As for reading, project students outgained controls at every grade level. A 2 × 5 analysis of variance yielded reliable effects for treatment ($F = 135.07$, $df = 1/522, p < .001$), for grade ($F = 20.76$, $df = 4/522$, $p < .001$), and for Grade × Treatment ($F = 9.00$, $df = 4/522, p < .01$). Specific comparisons between project and control means were reliable at each grade level. However, there was a much larger gain by project pupils in the sixth grade.

An analysis of covariance was performed on the year-end arithmetic scores of first-grade children with pretest IQ scores as the covariate. First-grade project children outperformed controls in mathematics ($F = 8.99$, $df = 1/62$, $p < .01$). The mean raw scores (adjusted scores are in parentheses) in arithmetic were 59.34 (57.39) for project

students and 43.83 (46.11) for controls. In terms of grade equivalent, first-grade project pupils scored at 1.70 grade level and controls at 1.40 grade level.

Discussion

The results of this study are promising. The authors believe a clear demonstration has been made that most, if not all, inner-city teachers can learn to use a positive contingency management procedure to insure behavior control, accelerated academic achievement, and probably as a function of the latter, substantial tested IQ gain. Further, it appears that teachers can maintain use of such a procedure over extended periods of time (two years for eight teachers in the present study) and that the technique works as well or better for children in the second year than in the first. In other words, there is substantial evidence that results of this contingency management procedure are enduring and not produced by some halo effect.

As Baer (1971) notes, research workers, including the present authors, have clearly shown that operant principles *can* be effectively applied in the classroom. However, previous investigators have not answered actuarial questions concerning the proportion of teachers that can be trained to use these principles or the proportion of children that can benefit from the program. The present study provides some preliminary answers to these questions. First, all 16 teachers participating in this project reduced their delivery of punishment and increased their delivery of reinforcement. At a subjective level, the authors, who visited these classes regularly, believe that all but 1 teacher learned to use these procedures effectively. Further, in all 16 classrooms, disruptive behavior dropped dramatically, and more significantly, task involvement increased. Finally, a simple count of the proportion of children showing any gain in reading from pre- to posttest indicated that 91% of the project children gained, whereas only 72% of the control children gained. Thus, contingency management appears to work well for most inner-city teachers and is effective with most inner-city children.

Informal observation also suggests many side benefits of the study, such as higher teacher morale, few if any disciplinary referrals by teachers to principals, and improved relations between school personnel and the community (as represented by the parents of children in the experimental classes).

The authors believe that the procedures of this study offer real hope for inner-city education and, they think, for the education of all children.[4]

REFERENCES

Baer, D. M. Behavior modification: You shouldn't. In E. A. Ramp & B. L. Hopkins (Eds.), *A new direction for education: Behavior analysis 1971.* Vol. 1. Lawrence: University of Kansas, Department of Human Development, 1971.

Becker, W. C., Engelmann, S., & Thomas, D. R. *Teaching: A course in applied psychology.* Chicago: Science Research Associates, 1971.

Becker, W. C., Madsen, C. H., Arnold, C. R., & Thomas, D. R. The contingent use of teacher attention and praise in reducing classroom behavior problems. *Journal of Special Education,* 1967, **1,** 287–307.

Bronfenbrenner, U. *Is early intervention effective?* Washington, D. C.: U. S. Department of Health, Education, and Welfare, Office of Child Development, 1974.

Clark, C. A., & Walberg, A. J. The influences of massive rewards on reading achievement in potential urban school dropouts. *American Educational Research Journal,* 1968, **5,** 305–510.

Coleman, J. S. et al. *Equality of educational opportunity.* Washington, D. C.: U. S. Government Printing Office, 1966.

Dittman, L. L. (Ed.) *Parent and child centers: A guide for the development of parent and child centers.* Washington, D. C.: Office of Economic Opportunity, 1967.

Hall, R. V., Lund, D., & Jackson, D. Effects of teacher attention on study behavior. *Journal of Applied Behavior Analysis,* 1968, **1,** 1–12.

Hamblin, R., Buckholdt, D., Ferritor, D., Kozloff, M., & Blackwell, L. *The humanization process.* New York: Wiley, 1971.

Harris, F. R., Wolf, M. M., & Baer, D. M. Effects

[4] The cost per pupil of this project was $375 per academic year including costs for research and project staff. If the latter two costs are removed, this program can be implemented for about $25 per pupil.

of adult social reinforcement on child behavior. *Young Children*, 1964, **20**, 9–17.

KUPERS, D. S., BECKER, W. C., & O'LEARY, K. D. How to make a token system fail. *Exceptional Children*, 1968, **35**, 101–109.

KVARACEUS, W. Negro self-concept. New York: McGraw-Hill, 1965.

MADSEN, C. H., JR., BECKER, W. C., & THOMAS, D. R. Rules, praise, and ignoring: Elements of elementary classroom control. *Journal of Applied Behavior Analysis*, 1968, **1**, 139–150.

McCANDLESS, B. R. *Children: Behavior and development*. New York: Holt, Rinehart & Winston, 1967.

McCANDLESS, B. R. *Adolescents: Behavior and development*. Hinsdale, Ill.: Dryden Press, 1970.

McLAUGHLIN, T. F., & MALABY, J. E. Development of procedures for classroom token economics. In E. A. Ramp & B. L. Hopkins (Eds.), *A new direction for education: Behavior analysis 1971*. Vol. 1. Lawrence: University of Kansas, Department of Human Development, 1971.

O'LEARY, K. D., BECKER, W. C., EVANS, M. B., & SAUDARGAS, R. A. A token reinforcement program in a public school: A replication and systematic analysis. *Journal of Applied Behavior Analysis*, 1969, **2**, 3–13.

STAATS, A. W. A case in and a strategy for the extension of learning principles to problems of human behavior. In A. W. Staats (Ed.), *Human learning*. New York: Holt, Rinehart & Winston, 1964.

STAATS, A. W., MINKE, K. A., FINLEY, J. R., WOLF, M., & BROOKS, L. O. A reinforcement system and experimental procedure for the laboratory study of reading acquisition. *Child Development*, 1964, **35**, 209–231.

STAATS, A. W., & STAATS, C. K. *Complex human behavior*. New York: Holt, Rinehart & Winston, 1963.

THOMAS, D. R., BECKER, W. C., & ARMSTRONG, M. Production and elimination of disruptive classroom behavior by systematically varying teacher's behavior. *Journal of Applied Behavior Analysis*, 1968, **1**, 35–45.

THOMPSON, M., BRASSELL, W., PERSONS, S., TUCKER, R., McCANDLESS, B. R., & ROLLINS, H. *Project Success Environment: A behavior modification program for inner-city teachers*. (Tech. Rep. APS-PSE-6 of Project Success Environment) Atlanta, Ga.: Atlanta Public Schools, 1973.

WOLF, N. K., GILES, D. K., & HALL, R. V. Experiments with token reinforcement in a remedial classroom. *Behavior Research and Therapy*, 1968, **6**, 51–64.

ZIMMERMAN, E. H., & ZIMMERMAN, J. The alteration of behavior in a special classroom situation. *Journal of the Experimental Analysis of Behavior*, 1962, **5**, 59–60.

(Received January 22, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 179-183

CONCRETIZATION AND SENTENCE LEARNING[1]

RICHARD C. ANDERSON[2]

Training Research Laboratory, University of Illinois, Urbana

One- to three-word modifiers that were judged to have the effect of making the subject noun phrase in each of a series of sentences more denotatively specific were prepared. In Experiment 1, people who received sentences containing concrete modifiers recalled significantly more words than people who received unmodified sentences, provided that in the former case the modifier was part of the retrieval cue. In Experiment 2, people exposed to concretely modified sentences recalled significantly more words than people who had seen redundantly modified sentences.

A striking fact is that the information embodied in concrete words is two or three times easier to learn than the information conveyed by abstract words (Paivio, 1969). Indeed, concreteness–abstractness, or "image evoking value," is the most potent determiner of the learnability of words that has yet been studied. The practical implications of this knowledge are not readily apparent, however. The teacher whose task is teaching about zebras can be encouraged to believe that his/her students will easily learn because "zebra" is a concrete term of high image-evoking value. But suppose he/she has the more difficult job of, for instance, teaching about regulations? What technique will increase the learnability of information to be acquired in connection with abstract terms such as "regulations?"

A plausible answer is to try to increase the concreteness of the abstract term by modifying it with concrete words. Consider this sentence:

The regulations annoyed the salesman.

Would it be learned more readily if modified as follows?

The strict parking regulations annoyed the salesman.

Answering the general form of this question was the main purpose of the experiments reported in this article.

Previous research on concretization has had mixed results. On the positive side, Yuille and Paivio (1969) and Montague and Carter (1973) have reported that vivid, concrete language facilitates learning from connected discourse. The present studies differed from these in one important respect: In the previous research, the concrete words with which the text was augmented became part of the to-be-recalled material, whereas in the experiments described in this report, the response was the same whether or not a concrete modifier was included. In other words, the present research aimed to show an indirect effect instead of a direct one.

On the negative side, Levin (1972) found that children learned somewhat fewer concretely modified nouns (e.g., spotted turtle) than unmodified nouns (e.g., turtle). However, Levin's experiments differed in many respects from the present ones. In the first place, his nouns were already denotatively specific; that is, they were specific enough so that the referent of each word could be pictured in a line drawing. No more than one or two of the (subject) nouns employed in the present studies could be represented unambiguously in pictures. In the second place, Levin employed the technique of free recall. Concreteness–abstractness does affect free recall, but the strongest effects are obtained when the concrete element serves as the cue in a cued-recall task, which was

[1] The author gratefully acknowledges the assistance of Valerie Koester, Claire Lieberman, Debra Sweet, Steven Sweet, and Peter Zych.

[2] Requests for reprints should be sent to Richard C. Anderson, Training Research Laboratory, 226 Education Building, University of Illinois, Urbana, Illinois 61801.

the arrangement in the studies reported herein. Furthermore, because free recall was required, subjects in Levin's study who received the concretized list had twice as many words to recall. This fact could explain his negative result.

A subsidiary purpose of Experiment 1 was to investigate the effects of variations in recall instructions. Recent research has indicated that people frequently substitute semantically related words for the verbatim language of the sentences they are trying to recall (Anderson, 1974). The question is whether recall instructions given *after* the sentences have been presented alter the frequency of substitutions. If the probability of a semantically related substitute word were to go down following verbatim recall instructions and go up following substance instructions, this would imply that some sentences or sentence constituents are coded in memory in both surface and semantic form. No difference in proportion of semantically related substitutions as a function of recall instructions would imply that there is a single memorial code for each verbal element.

Experiment 1

Method

Subjects. Involved in the experiment were 47 undergraduates, mostly women, for whom participation was a requirement in an introductory educational psychology course. Subjects were randomly assigned to conditions when they appeared for the experiment.

Materials. Sixteen simple declarative sentences in the past tense were constructed. Each had a concrete object noun and a general term for a subject noun. For each sentence, one- to three-word modifiers were prepared which were judged to have the effect of making the subject noun phrase more denotatively specific, that is, more concrete. Some examples of this were as follows:

The (ivory chess) set fell off the table.
The (huge earth-moving) vehicle missed the dog.
The (remote television) control pinched his hand.

Design. The two factors in the experiment were (*a*) type of sentence and retrieval cue and (*b*) type of recall instructions. One third of the subjects received both elaborated sentences and elaborated cues, one third received elaborated sentences but unelaborated cues, and the remaining third received both unelaborated sentences and unelaborated cues. "Elaborated" sentences and cues included the words modifying the subject noun, whereas "unelaborated" sentences and cues did not. Within each of the groups described thus far, half of the subjects received verbatim recall instructions and half received substance recall instructions. Verbatim instructions stressed literal reproduction of the sentences, whereas substance instructions indicated an answer would be counted correct if it contained the idea, or gist, of the sentence, whether or not identical words were used.

Procedure. Subjects were run individually in a small, sound-deadened cubicle. Learning instructions mentioned that a test was to follow but gave no details. The sentences, which were typed on 5 × 8 inch white unlined file cards, were presented at an eight-second rate paced by beeps from a tape recorder. Prior to each presentation, the cards were shuffled. After one exposure of the sentences, the subject solved addition, subtraction, and multiplication problems for 48 seconds to prevent recall from short-term and, probably, nonsemantic memory. Next came the recall instructions, either verbatim or substance. Finally, the test was presented. The subject noun or the subject noun phrase from each sentence served as the retrieval cue. The cues were typed on 4 × 6 inch white unlined file cards, which were shuffled before each use. The test was subject paced. The subject was instructed to give orally the entire sentence if possible but was also encouraged to give fragments when the rest of the sentence could not be remembered.

Scoring. The recall protocols were scored for numbers of verbs and objects recalled verbatim, discounting changes in number, tense, determiners, and auxiliaries, Also scored were the numbers of semantically related words substituted in place of the verbs and objects in the original sentences. A word was counted as semantically related if it was a synonym, close superordinate, close cohyponym, or hyponym of the word it replaced. These categories have been defined and illustrated elsewhere (Anderson, 1972, 1974). Several people scored the protocols. Disagreements were resolved in conference. Previous research has shown very high interrater agreement with respect to whether a word is semantically related or unrelated (Anderson, 1974).

Results

Table 1 contains mean proportions of verbatim, semantically related and total words recalled under the various conditions that prevailed in the experiment. The data were analyzed in a 3 × 2 (Type of Sentence-Cue Combination × Type of Recall Instruction) unweighted means analysis of variance. (Unweighted means were used be-

cause one cell was missing a case.) Type of sentence-cue combination had a significant effect when the dependent variable was number of words recalled verbatim ($F = 7.17$, $df = 2/41$, $p = .002$), number of semantically related substitutes recalled ($F = 9.76$, $df = 2/41$, $p = .000$), and total number of words recalled ($F = 12.40$, $df = 2/41$, $p = .000$). Neither type of recall instructions nor the Type of Recall × Sentence-Cue Combination interaction was significant in any analysis. An analysis in which the dependent variable was the arcsin transform of proportion of semantically related to total words recalled revealed no significant effects.

Newman-Keuls tests showed that the elaborated-sentence–elaborated-cue condition was significantly ($\alpha = .05$) better on all three measures than the elaborated-sentence–unelaborated-cue condition and significantly better on the semantically related words and total words measures than the unelaborated-sentence–unelaborated-cue condition. The elaborated-sentence–unelaborated-cue and unelaborated-sentence–unelaborated-cue conditions did not differ significantly on any measure. The detail of the data was generally consistent with the overall results. For instance, there were more total words recalled under the elaborated-sentence–elaborated-cue condition than under the unelaborated-sentence–unelaborated-cue condition for 15 out of the 16 sentences.

Of the 187 responses scored as semantically related substitutes for verbs or objects in the original sentences, 56% were scored as synonyms, 17% as superordinates, 15% as cohyponyms, and 12% as hyponyms.

Discussion

Experiment 1 gave rather clear evidence that concrete modifiers facilitate the learning of sentences whose subject nouns are general terms. Thus, the results were consistent with the original notions about concreteness and image-evoking value. But there is at least one other explanation of the data. Maybe *any* subject noun modifier would facilitate learning, perhaps by increasing orthographic or phonological distinctiveness. The purpose of Experiment 2 was to determine whether the functional attribute of the modifier was its effect on the denotative specificity of the subject noun phrase.

TABLE 1

MEAN PROPORTIONS OF WORDS RECALLED

Sentence cue	Words		
	Verbatim	Semantically related	Total
Elaborated–elaborated	.48	.21	.69
Elaborated–unelaborated	.26	.08	.34
Unelaborated–unelaborated	.38	.08	.46

Note. The total possible score was 32.

Subjects presented with elaborated sentences did poorly when they received unelaborated cues. Apparently, a vague cue was inadequate for the person to retrieve the representation which had been stored when the sentence was learned. Probably what happened was that the person tended to give a different interpretation to the general term alone than when it was concretely modified. This was unfortunate from a practical perspective for, to stretch a point beyond the data, the implication was that the concrete case employed to illuminate a generalization has to be reinstated if a person is to have access to his memory for that generalization.

It is quite interesting that people who received verbatim and substance recall instructions performed similarly on all measures including, especially, the proportion of semantically related to total words recalled ($F < 1.0$). This must mean that people do *not* have both a coding for the literal surface form of a sentence and a coding for its meaning. Otherwise, they would edit their production when given verbatim recall instructions so as to conform more closely to the original, whereas in all likelihood they would edit their output away from the original when given substance recall instructions so as to employ wording that seemed apt and tasteful. The data are consistent with the view that every sentence (which is learned at all) is represented in memory in

semantic form, though this theory is somewhat strained to explain why such a high proportion of words are recalled verbatim (see also Anderson, 1974). In this view, it has to be assumed that verbatim words appear in recall only when a person just happens to select the same lexical items when decoding the semantic representation into language. Another view consistent with the data is that some sentences or sentence constituents are coded literally, whereas the rest are coded semantically.

Experiment 2

Method

Subjects. The subjects were 28 undergraduates, mostly women, who participated to fulfill a requirement in an introductory educational psychology course.

Materials. Sixteen pairs of sentences were constructed. Within each pair, the sentences were identical except for one to three words which modified the subject noun. One of the sentences contained a concrete modifier of the same type employed in Experiment 1. The other entailed a redundant modifier, so called because it was judged to have little or no impact on the denotative specificity of the subject noun phrase. Consider the sentence,

The official regulations annoyed the salesman.

Most regulations are official. The class of official regulations is not much narrower than the class of all regulations. Other illustrations are as follows:

The oil-pressure gauge was covered with dust *versus* The measuring gauge was covered with dust.
The obscene exclamation embarrassed the nun *versus* The excited exclamation embarrassed the nun.
The sports periodical provided the information *versus* The regular periodical provided the information.

Design. The two factors in this experiment were (*a*) sentence list and (*b*) type of modifier. The same sentences, except for the modifiers, appeared in each list. Half of the sentences within each list contained concrete modifiers, half contained redundant modifiers. If a sentence included a concrete modifier in the first list, the parallel sentence in the second list included a redundant modifier. In other words, with respect to the factor of principal interest—type of modifier—this was a within-subjects, or mixed-list, design.

Procedure. Subjects were run individually in a small, sound-deadened cubicle. The sentences, which were typed on 5 × 8 inch white unlined file cards and randomized by shuffling before each use, were presented at an eight-second rate paced by beeps from a tape recorder. After one exposure to one of the two lists of sentences, the subject solved addition, subtraction, and multiplication problems for 48 seconds to prevent recall from short-term memory. Finally, the subject received the test. On each page of a 2¾ × 8½ inch answer booklet was mimeographed the subject noun phrase (always the entire phrase) of one of the sentences. The subject was instructed to write the rest of the sentence, all of it if possible, but any word or phrase he could recall if the whole sentence could not be remembered. Everyone received substance recall instructions. To reduce the likelihood of systematic position or sequence effects, the test booklets were collated in four different random orders and were assigned to subjects about equally often at random. The test was subject paced.

Scoring. The protocols were scored the same way as in Experiment 1.

Results

In Table 2 appear the mean proportions of words recalled. Analyses of variance showed that type of modifier had a significant effect on number of words recalled verbatim ($F = 15.81$, $df = 1/26$, $p = .001$), number of semantically related substitute words ($F = 14.36$, $df = 1/26$, $p = .001$), and total words recalled ($F = 27.87$, $df = 1/26$, $p = .000$).

List was nowhere a main effect (all Fs < 1.0), but in every analysis except the one involving total words, the List × Type of Modifier interaction was significant. These interactions occurred simply because the substitution of semantically related words was more likely in one list than in the other. Previous work has suggested that the probability of substitution is a function of the aptness of the wording of the original sentences and the availability of semantically

TABLE 2
Mean Proportions of Words Recalled

Modifier	Words		
	Verbatim	Semantically related	Total
Concrete	.34	.21	.55
Redundant	.22	.11	.33

Note. The total possible score was 16.

equivalent alternative wordings (Anderson, 1974). Apparently the two lists differed in these or other relevant respects.

Concrete modifiers led to better performance than redundant modifiers in 12 of the 16 sentences. There was little difference with the remaining 4 sentences. Of the 144 semantically related substitute words, 44% were scored as synonyms, 26% as superordinates, 17% as hyponyms, and 14% as cohyponyms.

General Discussion

The data show that concrete modifiers strongly facilitate the learning of sentences. The general educational implication for the teacher, author, and curriculum developer is to be as specific and concrete as possible. The results were positive enough to encourage research on other, more elaborate concretization techniques such as the use of metaphor, analogy, and physical models to represent systems of abstract concepts.

There was facilitation only when the entire, concretely modified subject–noun phrase served as the retrieval cue but not when the subject noun alone was the cue. Under the latter condition, it is quite likely that the subject noun was encoded differently at the time of original exposure and at the time of testing (cf. Martin, 1968; Wicker, 1970). It was clearly shown in Experiment 2 that the effective variable was concretization and not merely the presence of modifying words. Bower (1971) has mentioned, in passing, an unpublished experiment which also found no facilitation from redundant words on the stimulus side in a paired-associate task.

Throughout this paper the terms *concrete, specific,* and *vivid* have been used interchangeably. It should be noted that there is some indication that concreteness–abstractness and specificity–generality may be separate, though of course correlated, dimensions (Paivio, 1971, p. 83). Possibly vividness is also a distinct factor, though it appears to affect learning in the same way as concreteness/specificity. Kirchner (1969) inserted "vivid" or "dull" adjectives into appropriate and identical slots in a narrative. She found that people who heard the vivid narrative recalled more nouns than those who heard the dull narrative but, surprisingly, there was no difference in the recall of the adjectives themselves.

REFERENCES

Anderson, R. C. Semantic organization and retrieval of information from sentences. *Journal of Verbal Learning and Verbal Behavior,* 1972, **11,** 794–800.

Anderson, R. C. *Substance recall of sentences. Quarterly Journal of Experimental Psychology,* 1974, in press.

Bower, G. H. Mental imagery and associative learning. In L. Gregg (Ed.), *Cognition in learning and memory.* New York: Wiley, 1971.

Kirchner, E. P. Vividness of adjectives and the recall of meaningful verbal material. *Psychonomic Science,* 1969, **15,** 71–72.

Levin, J. R. *Verbal and visual processes in children's learning:* IV. *An attempt to reduce picture-word differences in free recall learning through concretization.* (Working paper 106) Madison: University of Wisconsin, Wisconsin Research and Development Center in Cognitive Learning, November 1972.

Martin, E. Stimulus meaningfulness and paired-associate transfer. *Psychological Review,* 1968, **75,** 421–441.

Montague, W. E., & Carter, J. F. Vividness of imagery in recalling connected discourse. *Journal of Educational Psychology,* 1973, **64,** 72–75.

Paivio, A. Mental imagery in associative learning and memory. *Psychological Review,* 1969, **76,** 241–263.

Paivio, A. *Imagery and verbal processes.* New York: Holt, Rinehart & Winston, 1971.

Wicker, F. W. On the locus of picture-word differences in paired associate learning. *Journal of Verbal Learning and Verbal Behavior,* 1970, **9,** 52–57.

Yuille, J. C., & Paivio, A. Abstractness and the recall of connected discourse. *Journal of Educational Psychology,* 1969, **82,** 467–471.

(Received April 2, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 184–188

SEVEN ASPECTS OF TEACHING CONCEPTS[1]

M. I. CHAS. E. WOODSON[2]

University of California, Berkeley

Seven steps (defining, identifying relevant attributes, identifying irrelevant attributes, listing exemplars, listing nonexemplars, describing the domain of the concept, and using analogies) in the instruction of concepts were isolated and their relative effectiveness examined by using each to teach verbal concepts (e.g., the "meaning" of Chinese characters) to college students. Subjects learned seven concepts at once in a within-subjects design. Instructions involving the definition or identification of the relevant attributes were found to be the most effective in terms of errors. Several kinds of generalization errors were isolated, and these were found to be associated with different instructional steps.

Most concept-learning experiments have involved displaying to the learner a collection of positive and negative instances from which he is to infer the concept being taught. On the other hand, when most of us set out to teach a concept to a learner in his natural habitat we attempt to take as much advantage as possible of the communication skills possessed by the instructor and the learner. This type of concept-learning task may be called the instructional paradigm to distinguish it from the reception paradigm (e.g. Haygood & Bourne, 1965), in which the instructor determines a sequence of positive and negative instances that are presented one at a time, and the selection paradigm (e.g., Bruner, Goodnow, & Austin, 1956), in which the learner selects which pattern comes next and all possible positive and negative instances are usually available.

Several instructional steps can be identified in the instructional paradigm situation. Among these are the (*a*) stating of a definition, (*b*) instructions intended to identify the relevant attributes of the concept, (*c*) instructions intended to identify the irrelevant attributes, (*d*) showing of examples of the concept, (*e*) showing of nonexamples of the concept, (*f*) description of the domain of the concept, and (*g*) use of analogies to describe the concept. This experiment was performed to compare these seven instructional steps in overall effectiveness in teaching concepts.

Evaluation of different instruction methods raises the question of amount of transfer. Classification practice may build up specific associations that appear to be concept learning on a classification task but lack the generality needed for a test of defining. Since a set to learn concepts facilitates concept learning (Reed, 1946), college subjects who are used to studying concepts may show even more transfer to testing tasks quite different from the instructional tasks. The question of transfer can be investigated by giving a variety of tests to all subjects.

This study proposes, therefore, to construct four measures of concept learning, to instruct concepts by seven alternative instructional methods within the instructional paradigm, and to compare achievements on all four tests following the instruction.

Method

Subjects

The subjects were 14 undergraduate students who responded to signs announcing a paid experiment. The data from two subjects were dropped

[1] This research was supported in part by a grant from the National Institute of Mental Health to the Institute of Human Learning.

[2] Requests for reprints should be sent to M. I. Chas. E. Woodson, Institute of Human Learning, University of California, Berkeley, California 94720.

from the analysis because they scored unusually high and admitted to having been coached by a previous participant in the experiment.

Design

Each subject learned seven concepts by one of seven methods of instruction. The concept–method pairs were counterbalanced so that among the 14 subjects each pairing occurred twice.

Task and Procedure

Subjects were instructed individually by instructions typed and drawn on IBM cards. Subjects were informed they would be paid 1¢ for each point scored on a test with 280 possible points over the material covered.

Seven concepts were taught at the same time. A unit of instruction consisted of a Chinese character and an English word drawn or typed on an IBM card. Each successive block of seven units contained one unit on each of the seven concepts. Within each block, the order was randomized independently for each subject. The rate of presentation was approximately one per 10 seconds. During the first two blocks, while the instruction card was in view of the learner, the experimenter read aloud the instruction appropriate to the instructional condition. After the two blocks involving treatment presentations, four instances and four noninstances, each labeled as an example or nonexample, were presented in the next eight blocks.

During training, no overt response was required of the subject, and the subject was instructed not to ask questions. After training there was a 10-minute interval before testing during which subjects filled out a survey that was a part of another experiment, and this interval was followed by a 28-item test on the concepts taught.

Concepts Taught

The concepts taught were "meanings" of the Chinese characters *wood*, *round*, *small*, *blue*, *hard*, *valuable*, and *eatable*. Sets of English words that involved a common concept were prepared as instances.

Training instances were English words that described things that were examples or nonexamples of the concepts being taught. For example, for the character *wood*, examples included lumber, tree, forest, board, baseball bat, and branch. Nonexamples included plastic, flowers, grass, purse, pliers, and apple.

Instructional Steps

The instructional steps used were as follows:

1. Definition: "This character refers to the round shape of objects."
2. Identification of relevant attributes: "The meaning of this character has something to do with the material of which things are made."
3. Identification of irrelevant attributes: "The meaning of this character has nothing to do with color."
4. Listing exemplars: "The following are things to which this character applies: most houses, most telephone poles."
5. Listing nonexemplars: "The following are things to which this character does not apply: fish, automobiles, and air."
6. Describing domain of concept: "This character applies to things which have a shape."
7. Use of analogies: "The meaning of this character (wood) is to this instance (tree), like metal is to an automobile."

Measures of Learning

Four measures of learning were used for each of the seven concepts. Three required written responses, while the fourth involved the classification of 10 not previously encountered instances as examples or nonexamples of the concept.

1. Definition. The subject was asked to "explain what the following character means."
2. Exemplars. The subject was asked to give several examples of things to which the character applies.
3. Nonexemplars. The subject was asked to give several examples of things to which the character does not apply.
4. Classification. The subject was asked to classify as examples or nonexamples a sample of 10 possible instances (5 examples, 5 nonexamples) that the subject had not encountered in training.

The order of the 28 test items (4 item types for each of the concepts) was randomly determined independently for each subject with the restriction that each block of 7 items contain all 7 concepts. The classification items were scored by counting the number classified correctly. The chance expectation of scores on this measure is about 5.

The definition, exemplar, and nonexemplar items required some judgment in scoring. They were scored by a person experienced in concept-learning studies, but unfamiliar with the experimental conditions involved. All answers to a particular type of item were displayed to him at once, and he was asked to classify each answer on an arbitrarily chosen 0–10 scale of relative effectiveness with a mean of approximately 5. His classifications were reviewed a number of times before the scoring was determined. The experimenter also examined the classifications and concluded there was close agreement between his and the rater's judgment.

Results

The concepts proved to be rather difficult, with a mean score of 5.3 out of 10 for the retention-transfer subtests. Table 1 gives the means across subjects and concepts for each of the instructional steps on the transfer-retention measures.

A multivariate analysis of variance,

TABLE 1

MEAN TRANSFER-RETENTION SCORE ON EACH OF THE MEASURES OF LEARNING FOR EACH INSTRUCTIONAL CONDITION

Instruction step	Measure of learning			
	Definition	Exemplar	Nonexemplar	Classification
Definition	7.7	7.6	9.0	8.0
Relevant attributes	8.4	7.8	8.6	8.6
Irrelevant attributes	3.1	1.7	4.7	5.0
Exemplars	7.4	8.0	5.3	6.5
Nonexemplars	3.0	1.3	7.9	5.1
Domain	6.2	7.5	6.5	7.2
Analogy	6.8	6.0	5.8	5.8

Concept × Treatment, with repeated measures on individuals and four measured variables, using Roy's largest root criteria (Morrison, 1967), indicated the seven instructional treatments differed ($\theta = .71$, $s = 4$, $m = 1/2$, $n = 34.5$, $p < .01$). Inferences about individual treatments and measures were made by ordering the means within measures and using individual t tests with $\alpha = .01$ distributed among all possible pairwise comparisons within each dimension by means of the Bonferroni inequality procedure (Miller, 1966).

The relative effectiveness of the several methods of instruction is summarized in Table 2. Methods of instruction are listed in order of effectiveness as measured by the retention tests, and methods that were not distinguished statistically are enclosed in brackets. Overall, the specification of the relevant attributes or a definition was most effective.

Examining the four measures of learning for each concept a subject learned, it is possible to classify the learning of each concept into a number of categories on the basis of the generalization apparent. These include: correct generalization in which the instances and noninstances are correctly classified; over generalization in which many noninstances are classified as instances; under generalization in which many instances are classified as noninstances; no generalization in which only the training instances and noninstances are recalled; and mis-generalization in which many instances and noninstances are misclassified. Table 3 summarizes the generalization observed under each instructional condition.

The classification of generalization errors used here was arrived at in a somewhat post hoc manner, that is, by examination of the outcomes and logical analysis considering these outcomes. In such a situation, statistical tests of hypotheses about the occurrence of these types of errors are not appropriate, as classifications arrived at may be due in part to the pattern of outcomes observed. It appears, however, from examination of the error frequencies, that certain generalization errors are associated with particular methods of instruction. For example, instruction about irrelevant attributes or nonexemplars appears to lead to over general-

TABLE 2

INSTRUCTIONAL CONDITIONS IN RANK ORDER ACCORDING TO EFFECTIVENESS FOR EACH MEASURE OF LEARNING

Measure of learning			
Definition	Exemplar	Nonexemplar	Classification
Relevant attributes	Exemplars	Definition	Relevant attributes
Definition	Relevant attributes	Relevant attributes	Definition
Exemplars	Definition	Nonexemplars	Domain
Analogy	Domain	Domain	Exemplars
Domain	Analogy	Analogy	Analogy
Irrelevant attributes	Irrelevant attributes	Exemplars	Nonexemplars
Nonexemplars	Nonexemplars	Irrelevant attributes	Irrelevant attributes

Note. Conditions in brackets are not distinguished statistically.

ization, while instruction about the domain appears to lead to under generalization.

Discussion

Much of the tradition in the study of concept learning has been restricted to the study of Weigl-type problems with well-defined stimulus dimensions. Hovland (1952) and Hovland and Weiss (1953) argued that it was impossible to evaluate the information provided by instruction if the task did not involve a known number of attributes. This analysis has inspired a large number of studies using stimulus patterns with a small number of discrete and obvious levels, studies that are referred to as concept learning but whose generalizability and applicability to human concept learning in the natural habitat are very suspect. The emphasis of these studies has unfortunately been upon what the experimenter could detect was included in the instruction, but the critical factor is what the learner can detect and thereby exhibit by his behavior. In other words, instruction is not the exhibiting of information, but it is transmission to *and* reception by the learner. Reception is the essential characteristic, not display.

Furthermore, the question is not really one of whether the experimenter can analyze the information content of the task, but whether the learner can do so in the instructional situation. If the information content of the task is easily analyzed, learners may well use different learning strategies than if the information content is not open to immediate analysis. Differences in the information communicated by instructional methods need not be measured to compare methods. Differences among instructional methods, as measured by what the learner can do, are the appropriate measure of instructional effectiveness.

In the present study, the instructional steps were probably not equivalent in terms of information provided, but the approach was to avoid concern with equating instructional methods in terms of the amount of information available in the eyes of the experimenter, to equate instruction in terms of time and effort on the part of the learner, and to consider any differences in amount of information communicated, as measured by what the learner achieves, the basis of comparing instructional methods.

TABLE 3

Inferences Regarding the Generalization Observed in Each Instructional Condition, Based on an Examination of All Four Measures of Learning

Instruction step	Correct	Over	Under	None	Miss
Definition	11	1	1	0	1
Relevant attributes	9	0	2	2	1
Irrelevant attributes	0	9	2	1	4
Exemplars	6	1	1	4	2
Nonexemplars	0	7	2	0	5
Domain	7	1	5	0	1
Analogy	3	6	1	1	3

The results indicated the seven instructional steps were of different effectiveness and that these differences were detectable by relatively simple tests after a short period of instruction. It may be that the learning of several concepts at once accentuated these differences as compared to other experiments in which only one concept at a time was learned.

Since the subjects in this study were college students, it was reasonable to infer that they were rather experienced concept learners. It may be that the relative importance of these instructional acts depends somewhat upon the subjects' experience in dealing with them. Such a situation was found in the case of dealing with negative instances.

The measure of learning does seem to be related to the method of instruction, although the differences are not as striking as might be expected. Transfer from one method of instruction to dissimilar testing procedures may be expected due to the subjects being highly practiced concept learners who have received much training intended to facilitate transfer.

These results imply that the identification of the relevant attributes (average rank of 1.5 among the methods used, see Table 2) is the most effective instructional strategy, followed closely by definition (average rank of 2.125 among the methods used). This

may be due in part to the character of the task used, which may have been primarily a concept-identification task for most subjects.

This study suggests a number of questions that are unanswered by the present data. Do persons differ as to which instructional methods are more effective? Are these differences great enough to make individual instructional prescriptions beneficial? What combinations of instructional methods are the most effective in terms of learner and instructor, time and effort? Are the differences in the instructional effectiveness among methods due to experience of the subject or the characteristics of the methods themselves? In the face of detected errors of generalization, which method of instruction is most effective? The method associated with low numbers of generalization errors of that particular type may be the most effective for remedial instruction. The within-subjects design used in this study may well not be the most effective way of going about answering these questions.

REFERENCES

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. *A study of thinking.* New York: Wiley, 1956.

HAYGOOD, R. C., & BOURNE, L. E., JR. Attribute- and rule-learning aspects of conceptual behavior. *Psychological Review,* 1965, **72,** 175–195.

HOVLAND, C. I. A "communication analysis" of concept learning. *Psychological Review.* 1952, **59,** 461–472.

HOVLAND, C. I., & WEISS, W. Transmission of information concerning concepts through positive and negative instances. *Journal of Experimental Psychology,* 1953, **43,** 175–182.

MILLER, R. G. *Simultaneous statistical inference.* New York: McGraw-Hill, 1966.

MORRISON, D. F. *Multivariate statistical methods.* New York: McGraw-Hill, 1967.

REED, H. B. Factors influencing the learning and retention of concepts: I. The influence of set. *Journal of Experimental Psychology,* 1946, **36,** 71–87.

(Received January 23, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 189-201

TEACHING FACTS ABOUT DRUGS:
PUSHING OR PREVENTING?[1]

RICHARD B. STUART[2]

University of Michigan

Nine hundred thirty-five seventh- and ninth-grade students in two suburban junior high schools were randomly assigned to experimental drug education or control groups. A 10-session fact-oriented drug education program was offered in two formats (student or teacher led) and with three sets of contents (lesser drugs only, major drugs only, or both sets combined). The program was evaluated through the use of a self-report measure of drug information, drug use, and attitudes relating to drug use. Results indicated that relative to controls, subjects receiving drug education significantly increased their knowledge about drugs, their use of alcohol, marijuana, and LSD, and their sale of the latter two drugs, while their worry about drugs decreased. Neither format nor content factors were shown to influence the results of the program. When the interaction among drug use, knowledge, and worry was examined, it was shown that use tends to rise as a function of the combination of increased knowledge and reduced worry. This combination of factors was not sufficient as a predictor of drug use, however, suggesting the influence of other, untested factors. Within the limitations posed by several qualifications, it is suggested that these findings support the notion that drug education may not necessarily be positive in its effect, indicating the need for precise measurement of program outcomes.

There is a growing belief that the use of some drugs such as alcohol and soft hallucinogens is increasing at all strata of society. In reponse to this, some jurisdictions have sought to control drug use through stricter law enforcement, while others have responded by decriminalizing the possession and/or use of some drugs (Stachnik, 1972). Whatever the legal response to the presumed rise in drug use, there has been a widespread increase in reliance upon drug education as a preventive measure. At least 24 states now require drug education in the public schools (National Commission on Marijuana and Drug Abuse, 1972), with states such as Michigan requiring "education in health related topics with special reference to the nature of tobacco, alcohol and narcotics and their effect on the human system [*Michigan State School Code*, 1955]."

Drug education has taken many forms involving varying permutations of settings, goals, targets, methods, educators, and contents. Settings have ranged from reliance upon mass media through formalized programs in school and work institutions to individualized counseling in drug crisis centers. The goals set typically call for either the elimination of all drug use or, failing this, the reduction of use to levels

[1] This research was sponsored by grants from the Office of Drug Abuse and Alcoholism, State of Michigan, and the Board of Education, Ann Arbor, Michigan.

The author wishes to acknowledge the significant contributions to this research made by Sharon Schafer, who both served as a teacher and who coordinated much of the data collection; Ann Patton, who served as a teacher and significantly contributed to the development of the educational program; Patricia Ciriello, who coordinated the coding of the bulk of the data; Kenneth Guire, who planned much of the data analysis; and Lynn Nilles, who both administered the research and edited the final manuscript.

[2] Requests for reprints, copies of the curricular outline, and the measurement instrument used in this research should be sent to Richard B. Stuart, who is also at Behavior Change Laboratories, 3156 Dolph Drive, Ann Arbor, Michigan 48103.

that minimize physical or emotional health hazards. The targets have been nonusers who are believed to be potential users, experimental users, committed users, or the parents, teachers, employees, or friends of members of any of these groups. The methods used have included fear induction, exhortation, authority- and entertainment-based appeals, role playing, encounter groups, and formal lectures (Richards, 1970). The educators have been drawn from throughout the ranks of the professional and lay communities, from former drug users and from among the target populations themselves. Finally, the contents have ranged from reliance upon unbiased factual presentation of drug facts through the discussion of the social and legal ramifications of drug use to the introduction of materials that reflect a very strong antidrug bias.

With such a variety of programs, it is reasonable to expect that some programs have more positive results than others. As an optimal outcome, the programs might succeed in eliminating or temporizing drug use. As a midrange outcome, the programs might change attitudes or the level of drug information, which has no direct bearing upon drug use. As a negative outcome, the programs might exacerbate drug use by (*a*) providing students with sufficient information to facilitate the initiation of use; (*b*) providing students with facts that overcome the prejudices that had been inhibiting use; (*c*) desensitizing students about drugs through repeated discussion of drug concepts in environments such as schools, which have been traditionally disassociated from drug use; (*d*) leading students to think of themselves as potential drug users merely by virtue of their having been included in drug education programs; (*e*) changing attitudes that were the bastion of defense against drug use; or (*f*) occasionally including inaccurate or biased information, which undermines the credibility of the basic educational message.

The risk of negative effects of drug education is suggested by evidence that shows that a relatively high level of knowledge about drugs is associated with higher levels of drug use. In a survey of adolescents in four Michigan communities, Stuart and Schuman (1972; see also Swisher, Crawford Goldstein, & Yura, 1971) reported:

> The non-users of every type of drug were found to have lower drug information scores than did the users, the difference being statistically significant for every drug except alcohol . . . . Further, there was a strong relationship between past frequency of use of a drug and drug information scores for all drugs except alcohol [Stuart & Schuman, 1972, p. 139].

While these correlational findings do not demonstrate that increases in drug information necessarily cause or even catalyze increases in drug use—for they may follow rather than precede the initiation or intensification of use—they do show that drug information per se does not necessarily inhibit use.

Because of the possibility that drug education may actually be associated with intensified use, it is particularly unfortunate that the surfeit of drug education efforts has been associated with a paucity of program evaluations. Few states require any evaluation at all. In its exhaustive search, the National Commission on Marijuana and Drug Abuse (1972) did not report one evaluation procedure that adequately accounted for the effects of the countless programs that it cited (see also Ford Foundation, 1972). Those few states that do require any evaluation at all tend to be concerned with issues of secondary importance such as "the question of the age at which instruction should begin, how often programs should be repeated, and whether programs should be taught by medical and law enforcement personnel . . .. [California State Assembly, 1967]." When use of drugs has been used as a dependent variable (e.g., California Department of Education, 1970), the consequences of drug education have not been encouraging. Nevertheless, there is little official questioning of the desirability of drug education, with the focus of the sparse evaluations likely to be placed on matters of secondary importance such as consumer appeal.

The present study is an effort to measure the primary effects of drug education—its impact upon use of drugs and attitudes relating to use—as well as evaluating the effects of differing instructional patterns.

The goal of this program was the prevention or reduction of use among the junior-high-school-age targets. The settings were two junior high schools in an upper-middle-class suburban university community. The methods included lectures by teachers and student presentations designed to communicate facts about the physiology and pharmacology of drug use in association with its legal, social, and psychological ramifications. The two teachers were women in their mid-twenties, one of whom was a registered pharmacist with extensive experience in counseling committed drug users, while the other was an experienced junior and senior high school teacher who had worked in inner-city schools. Both were casual in appearance and relaxed in style. Three content divisions were used in this study. Some students received education about the lesser drugs—alcohol, tranquilizers such as Valium, marijuana, hashish, nicotine, and caffeine (Type A content); others received education about the major drugs—LSD and other hard hallucinogens, amphetamines, barbiturates, and narcotics (Type B content); while still others received education about all of the drugs (Type AB content). All data were collected by the two teachers, whether from experimental or control subjects.

The dependent variables in this study were measured by a single instrument that assessed knowledge about drugs, self-report of present and past use and the sale of drugs, and three attitudes believed to be associated with drug use—worry about drug effects, acceptance of drug use as a nondeviant action, and alienation from more conventional sources of satisfaction.

Method

Subjects

The subjects were 935 junior high school boys (55.5%) and girls (44.5%) of whom 63.5% served as experimental subjects and 36.5% served as controls. Five hundred and nine seventh graders were drawn from science and unified studies classes and 426 ninth graders were drawn from civics classes. These required classes were chosen because they were formed on an alphabetical rather than an ability–interest basis, thereby affording a representative cross section of the seventh- and ninth-grade students in the school. These classes were then randomly assigned to either the experimental or control conditions, and parents were duly notified of the assignment of their children. Fewer than 1% of the parents took advantage of the option to have their children withdrawn from the program prior to its inception, and only a small number of students were withdrawn after the program began. The first-semester classes provided pre- and post-experimental data and follow-up data, and the second-semester classes provided pre- and post-experimental data only. All data were collected during October, January, and April of the 1971–1972 academic year.

Instrument

A single instrument measuring three clusters of factors provided the data for this investigation. The first portion of the instrument asks for self-report of the present and past use of seven classes of drugs: alcohol, soft hallucinogens, hard hallucinogens, stimulants, depressants, narcotics, and solvents. For example, in measuring present use, respondents are asked to indicate whether they do not use a particular drug, use it once or twice a year, once or twice a month, once or twice a week, or daily. These answers can then be quantified simply by assigning numerical values from 1 (nonuse) to 5 (daily use). While not precisely identical to the frequencies stated in the question, this procedure nevertheless yields an approximation of use via an ordinal scale. Furthermore, this scale is conservatively biased inasmuch as the transitions at the lower end of the scale (e.g., from nonuse to use once or twice per year) are of a lesser order than the transitions at the upper end of the scale (e.g., from weekly to daily use).

These questions were previously used by Stuart and Schuman (1972) with apparently good results. The reliability of the questions was established in two ways. First, the questions were administered to 50 subjects within a seven-day period and yielded a test–retest reliability of .86. Second, scores of the two questions containing "do not use this drug" as an alternative and of the four questions containing "have not used this drug" as an alternative could be cross-checked for each of the seven classes of drugs investigated. The first set of questions yielded 7 comparisons (one for each class of drugs), while the second set of questions yielded 42 comparisons (all six permutations of the four questions for each of the seven classes of drugs). The average level of consistency in answering these questions (i.e., the number of subjects giving the same answer to each pair of questions) ranged from 97.3% to 99.8%, reflecting a high level of reliability.

The face validity of these questions could be examined only in terms of the conformity to a priori expectation of answers to varying questions. The six tests of face validity were as follows:

1. It was shown that the frequency of use of drugs at least once declined steadily from 81.2% for alcohol to 1.3% for narcotics, with virtually no reported use of solvents.

2. No subject who reported a high level of present use of a drug (e.g., monthly, weekly, or daily use) did not also report a high level of past use of that drug (e.g., 10 or more times.)

3. The percentage of subjects reported having sold marijuana was 7.4% as opposed to 3.3%, who reported having sold LSD.

4. The number of subjects who reported having injected LSD, stimulants, depressants, narcotics, or solvents was under 1.5%.

5. The number of bad drug experiences reported corresponded with the number of times in which subjects reported having used drugs in the past (i.e., for alcohol, marijuana, and LSD, the correlations were .32, .35, and .36, respectively [$df = 173$, $p < .01$]).

6. The frequencies of use of drugs at least once, reported in this study, closely corresponded to those found in a survey conducted in the same community one year earlier (Stuart & Schuman, 1972); for example, both studies found that approximately 74% of the seventh and ninth graders had used alcohol at least once, while marijuana and LSD were used by approximately 23% and 8% of these same groups, with use levels being slightly lower in the present investigation.

Taken together these findings appear to bear out the face validity of these questions, with predictive validation depending upon data that were unavailable in this study (i.e., the number of police contacts, peer reports of drug use, or contacts with drug crisis treatment centers).

The second portion of the instrument contains 14 items measuring worry about drugs (e.g., "I am afraid of becoming addicted"), 5 items measuring the span of drug-related deviance (e.g., "As long as the buyer wants drugs, there is no reason to feel guilty about selling them"), and 5 items measuring drug-related alienation (e.g., "People use drugs because they find it hard to be happy other ways"). Each item was answered on a 5-point scale ranging from strongly agree to strongly disagree. Low scores indicate low worry and deviance tolerance, while high scores indicate a high level of alienation. The Kuder-Richardson reliabilities for each of these scales were .92, .80, and .79, respectively, with corresponding test–retest reliabilities of .88, .81, and .61. Responses to the three scales were also significantly intercorrelated among the 757 subjects, whose responses to all three scales were complete at the time of the pretest. These correlational coefficients were −.43 for worry and deviance, .51 for deviance and alienation, and −.54 for worry and alienation ($p < .001$). Thus, as expected, those who worried the most about drugs had the lowest deviance scores and those who felt themselves to be the most alienated had the highest deviance scores. (No a priori expectations were entertained for the association between worry and alienation.)

Four partial validity checks were undertaken. In the first of these, the scores of each of the scales were intercorrelated with the subjects' reports of their use of alcohol, marijuana, and LSD. These correlations were as follows:

	Alcohol	Marijuana	LSD
Worry	−.34	−.44	−.22
Deviance	−.32	−.48	−.33
Alienation	.20	.14	.15.

All of these correlations were significant ($df = 781$, $p < .05$) and in the expected direction with the correlations for alienation being quite low. In the second validity check, the deviance scale responses of those who did and did not report that they had sold marijuana were compared. The deviance scale had a range of 5 (a high deviance score) to 25 (a low deviance score). The 82 subjects who reported having sold marijuana had a mean score of 12.51 ($SD = 13.39$) indicating higher deviance, while the 774 subjects who reported not having sold marijuana had a mean score of 15.506 ($SD = 12.59$), indicating lower deviance. A test of these differences yielded a t of 7.2441 ($df = 854$, $p < .001$). Third, the deviance scores of subjects who reported that they could obtain alcohol, marijuana, or LSD very easily, easily, or with difficulty or not at all were compared. All three analyses of variance yielded significant results ($F = 5.0957$, $df = 511$, $p < .001$; $F = 9.6136$, $df = 505$, $p < .001$; $F = 9.9719$, $df = 493$, $p < .001$, for the three drugs, respectively). In each instance, deviance tolerance was higher for subjects finding it easy rather than difficult to obtain drugs. For example, the average deviance score of those reporting ease in obtaining LSD was 13.40, as opposed to 15.08 for those reporting great difficulty in obtaining the drug. Finally, the extent of worry about drugs was intercorrelated with the number of bad experiences at the level of .23, a low but significant ($p < .05$) correlation. In view of these four sets of findings, it can be concluded that the three sets of attitude measures have at least defensible fact validity.

The third portion of the instrument measured the subject's knowledge of the pharmacology, psychological effects, and the legal implications of drug use. Items in this series were multiple multiple-choice questions, for example, "An overdose of which of the following can cause death? Alcohol? Methedrine? Morphine? Seconal?" or "Drugs are metabolized (broken down) in the: small intestine? liver? kidneys? pancreas?" This 78-item scale was also used by Stuart and Schuman (1972), where its Kuder-Richardson reliability attained a level of .96 and a test–retest reliability of .85.

In addition to these three substantive areas, the instrument also contained questions needed for sample descriptions such as grade, sex, school, etc. Finally, in keeping with the suggestion of King (1970), all questionnaires were completed anonymously and were identified for matching purposes only by an alphanumeric code: the last three letters of the respondent's mother's maiden name and the last two digits of the respondent's house number. Because only a small number of these protocols could be matched for pre- and post-test data (594) and fewer still for pretest, posttest, and follow-up data, the analyses reported here

are posed exclusively upon the pooled group responses of the full 935-subject sample.[3]

Procedure

Drug education was offered one day a week in each participating class for a period of 10 weeks. Most individual sessions included a pretest and posttest of knowledge of the topics under discussion and were built around a combination of lecture, discussion, and audiovisual techniques. The topics for each of the sessions were as follows: pharmacology and physiology background, drugs and the nervous system, drugs and society, and discussions of each of the seven classes of drugs identified in the questionnaire. During the first semester, one half of the experimental classes were randomly assigned to either teacher- or student-led patterns of instruction. Both groups had the same three teacher-led general presentations (e.g., drugs and society). However, while the teachers continued to lead the instruction about particular drugs in half of the classes, in the other half of the classes students met in groups to develop educational materials and then presented these materials to the class. Because attendance and class participation levels were lower for student-led instruction and because there were no significant outcome differences between teacher- and student-led groups after the first semester (see *Data Analysis*), both were combined for the final data analysis, and all instruction during the second semester was teacher led. During both semesters, the content of instruction was divided in a manner that permitted one third of the experimental subjects to receive the A, B, or AB program contents.

Data Analysis

Data were analyzed using conventional analysis of variance, covariance, *t*-test, and chi-square programs available through the MIDAS system of the Statistical Research Laboratory, University of Michigan. To evaluate the relative contributions of knowledge and worry to the use of drugs, the Kullbach information or likelihood ratio was used (Kullbach, 1968, pp. 159–169; Kullbach, Kupperman, & Ku, 1962). This procedure permits a three-way analysis of (*a*) the independence of three variables, (*b*) the extent to which column variables are independent of layer variables, (*c*) the extent to which row variables are independent of layer and column variables, (*d*) the extent to which row variables are independent of layer variables, and (*e*) the extent to which row variables are independent of column variables for each level of the layer variable. If drug use were the column variable, worry about drugs, the row variable, and level of knowledge, the layer variable, the information ratio would indicate the level of independence of variables at all points of intersection. The information ratio thus yields a higher level of analysis than would be available through the use of chi-square, yet its use is appropriate whenever chi-square can be applied.

In the analysis that follows, data are presented for only three classes of drugs: alcohol, marijuana, and LSD. The remaining classes of drugs (stimulants, depressants, narcotics, and solvents) were used by fewer than 5% of the subjects at all times of testing and, therefore, did not yield sufficient variance to permit an analysis of group differences. Finally, all analyses in this report were based upon the number of subjects who completely answered the relevant question. Because it was not uncommon for subjects to omit one or more sections of any given question, the number of subjects contributing to the various analyses was quite variable, although the number of protocols with missing data was quite consistent with the expectation for the present type of research and subject population.

Results

Drug Education

The major results of this study are summarized in Table 1 in which it can be seen that at pretest, seventh graders had predictably less information about drugs than did ninth graders, and experimental subjects at both grade levels had more drug knowledge than did controls at similar levels. Also, experimental subjects gained more knowledge than did controls, and the differences between experimentals and controls narrowed slightly but were nevertheless sustained at follow-up. Similar patterns were found for drug use and worry about drugs.

For each of the variables in Table 1, an analysis of variance was performed at pretest, posttest, and follow-up times to determine the significance of the three-way interaction between condition (experimental versus control), grade (seventh versus ninth)

[3] The results presented are all derived from analysis of group comparisons. Because the questionnaires of only approximately one out of five (first-semester) students could be matched on pretest, posttest, and follow-up, no analyses were performed using these data. Analyses of covariance were performed for the four out of nine subjects whose pretest and posttest questionnaires could be matched. These analyses all closely paralleled the results from the full subject cohort with two exceptions: As opposed to the larger group, change in worry about drugs was marginally nonsignificant ($p < .06$), and the change in deviance was not significant at posttest. To avoid redundancy these data are not reported here, but tables are available upon request.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF SEVENTH- AND NINTH-GRADE SUBJECTS ON USE OF ALCOHOL, MARIJUANA, AND LSD AND ON SCALES OF KNOWLEDGE AND WORRY ABOUT DRUGS

Source	Seventh grade									Ninth grade								
	Pretest			Posttest			Follow-up			Pretest			Posttest			Follow-up		
	$\bar{X}$	*SD*	*n*	$\bar{X}$	*SD*	*n*	$\bar{X}$	*SD*	*n*	$\bar{X}$	*SD*	*n*	$\bar{X}$	*SD*	*n*	$\bar{X}$	*SD*	*n*
Drug knowledge																		
Experimental	36.49	9.56	278	44.11	9.93	273	41.58	9.20	103	39.40	10.35	241	47.73	9.57	218	44.41	9.94	80
Control	35.36	9.38	155	37.86	9.96	169	36.21	9.45	80	38.29	10.89	147	40.76	10.81	139	39.65	9.97	72
Use of alcohol																		
Experimental	1.96	1.02	324	2.26	1.10	277	2.38	1.21	115	2.40	1.14	263	2.60	1.14	236	2.74	1.14	94
Control	1.82	.92	183	1.82	.98	182	1.93	.99	89	2.33	1.10	163	2.40	1.07	169	2.37	1.14	78
Use of marijuana																		
Experimental	1.11	.52	323	1.24	.74	279	1.51	.93	115	1.58	1.24	263	2.02	1.38	234	2.33	1.47	92
Control	1.05	.28	185	1.12	.50	182	1.10	.54	89	1.55	1.09	160	1.57	1.11	166	1.69	1.22	78
Use of LSD																		
Experimental	1.01	.08	324	1.09	.54	277	1.23	.64	116	1.18	.55	262	1.35	.71	234	1.42	.98	95
Control	1.03	.31	185	1.03	.24	184	1.02	.21	90	1.10	.42	158	1.14	.51	167	1.21	.74	78
Worry about drugs																		
Experimental	16.17	3.33	317	13.57	3.76	278	11.95	4.16	115	14.63	3.51	260	12.40	3.62	231	10.88	3.99	93
Control	16.22	3.64	185	15.86	3.89	186	15.66	3.51	91	14.49	3.87	163	15.11	3.72	162	15.91	3.12	74
Deviance																		
Experimental	15.93	3.02	312	15.68	4.07	276	15.45	4.16	115	14.71	3.38	264	14.40	3.62	231	13.88	3.99	93
Control	16.07	3.81	179	15.74	3.89	185	15.55	3.49	90	15.12	3.35	162	15.11	3.72	162	15.90	3.12	74
Alienation																		
Experimental	13.64	3.38	316	13.65	3.84	277	13.48	4.29	115	13.87	3.39	264	13.86	3.50	230	14.01	4.36	93
Control	13.72	4.00	185	14.12	3.77	186	14.03	3.79	91	13.96	3.15	161	13.88	3.79	162	14.46	3.15	74

and sex. For the first five variables, all 18 F ratios reached the level of significance ($p < .05$). Independent a posteriori t tests were performed to locate specific significant differences between the means of each of these groups singly and in interaction with each other. This analysis revealed that experimental subjects did not differ significantly from controls before undergoing the educational program, but they did show significantly ($p < .05$) more knowledge, greater alcohol, marijuana, and LSD use, and less worry than did controls at posttest and at follow-up. Through t tests it was also shown that seventh- and ninth-grade subjects differed in each of these responses at all three testing times except for the follow-up assessment of drug information, when ninth-grade subjects scored higher, although not significantly so.

The only significant differences not involving the experimental condition, grade, or their interaction were sex differences found in the lower alcohol use reported by girls at pretest and the greater worry reported by girls at follow-up. In a marginally significant three-way interaction, seventh-grade experimental girls were found to have higher knowledge than did their male peers, while ninth-grade girls outperformed boys on the knowledge inventory in both experimental and control conditions, although not significantly.

In the final two areas covered in Table 1, measures of deviance tolerance and alienation, a less definitive pattern of results emerged. Significant F ratios were obtained for the deviance measure at pre- and posttest, but the F ratio at follow-up was not significant. At pretest, significant t tests showed ninth graders to be more accepting of drug use as nondeviant than were seventh graders ($p < .001$) and boys to be more accepting than were girls ($p < .001$). At posttest, ninth graders continued to be more accepting than were seventh graders ($p < .001$), but the sex difference was replaced by a significant increase in the acceptance of drugs by experimental subjects as opposed to controls ($p < .01$). None of these differences were maintained at the follow-up, however. While it is possible that the program had merely a short-lived impact upon this attitude, it is more likely that the mid-semester passage of a local ordinance making marijuana possession a $5 misdemeanor undoubtedly lessened the extent to which both experimental and control subjects considered marijuana use to be deviant. Because the alienation scale yielded no significant F-ratio differences, further analysis of data generated by it was not undertaken.

Figure 1 presents changes in the percentage of experimental and control subjects reporting that they had sold either marijuana or LSD before and after the educational period and at follow-up. The figure clearly demonstrates that more experimental than control subjects reported having sold drugs at pretest, with the differences accelerating at posttest for both drugs and at follow-up for marijuana only. Chi-square tests of the significance of these differences were computed for both drugs at each time of testing.

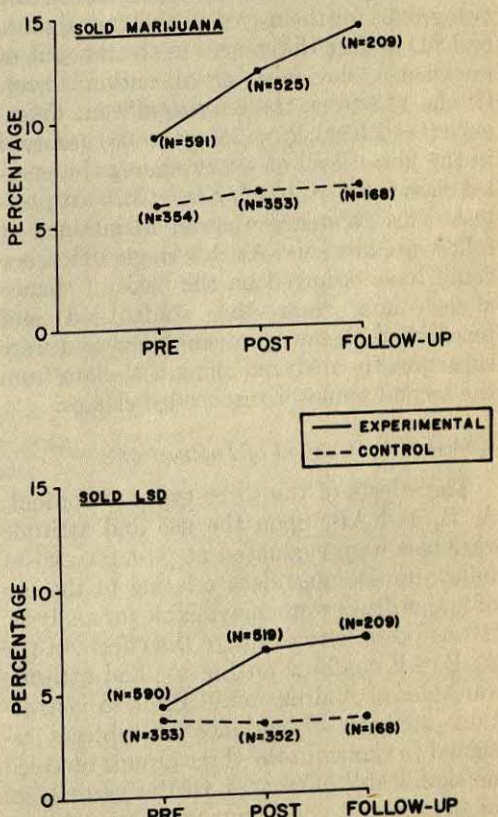


FIGURE 1. Percentages of students who sold marijuana and LSD at pretest, posttest, and follow-up.

The pretest differences for marijuana and LSD were both nonsignificant ($\chi^2 = 3.1121$ and .1756, respectively, $df = 1$), while posttest differences were significant for both drugs ($\chi^2 = 7.6635$ and 5.0276, respectively, $df = 1$, $p < .03$). At follow-up, there were significantly more marijuana sellers in the experimental group ($\chi^2 = 5.8557$, $df = 1$, $p < .02$), but the difference in the number of sellers of LSD was not significant ($\chi^2 = 2.2615$, $df = 1$).

Effects of the Type of Instruction

Differences between the effects of the student-led instruction received by 128 students and the teacher-led instruction received by 136 students during the first semester were evaluated by t tests. There were no significant differences at pretest, estabilishing the initial equivalence of the groups. At posttest and at follow-up there were no significant differences between the two groups for the use of alcohol, marijuana, or LSD and no differences in the amount of knowledge, deviance, or alienation scores. Of the 14 t tests, the only significant difference ($t = 2.0540$, $df = 230$, $p < .05$) occurred in the lower level of worry among the peer-led classes ($\bar{X} = 14.09$, $SD = 3.76$) at posttest. This difference was not maintained at follow-up, however. As this single difference could have occurred on the basis of chance alone, data from the student-led and teacher-led classes were combined and were subsequently analyzed along with data from the second-semester teacher-led classes.

Effects of the Content of Instruction

The effects of the three types of content, A, B, and AB, upon the use and attitude variables were evaluated at posttest and at follow-up. Because data relating to the use of major drugs were unavailable for analysis, it was possible to evaluate the effects of the A, B, AB contents on the use and attitude variables of A drugs only. Prior to instruction, analyses of variance of subjects assigned to classes in the three groups revealed no significant differences. Of the 14 analyses of the posttest and follow-up data, only one was significant. There was a significant difference for alcohol use at follow-up ($F = 4.2647$, $df = 190$, $p < .02$) in which it was found that those receiving AB content used significantly more alcohol ($\bar{X} = 2.87$, $SD = 1.24$, $n = 63$) than did those receiving only B content ($\bar{X} = 2.27$, $SD = 1.45$, $n = 70$) but not significantly more than did those receiving only A content ($\bar{X} = 2.50$, $SD = 1.58$, $n = 58$). Because this single finding might have occurred by chance, it has to be concluded that the content of the drug curriculum tended to be nonspecific in its effects; that is, changes in drug-use patterns and attitudes about use tended to occur independently of the specific content of instruction.

Concomitants of Use

In order to explore some of the concomitants of subjects' decisions to initiate experimentation with drugs, the impact of knowledge and worry were explored. First it was found by chi-square analysis that when knowledge and worry were each divided into high, medium, and low categories, their interrelationship was not significant ($\chi^2 = 8.3652$, $df = 4$, $p < .10$). Tables 2 and 3 show the knowledge and worry scores of subjects reporting use of alcohol, marijuana, and LSD. It can be seen that for marijuana only a significantly higher percentage of users as opposed to nonusers is found to have high knowledge, while significantly more users of all three drugs had less worry than nonusers. It is also obvious, however, that from one-fifth to one-third of the users of all three drugs were in the lowest knowledge category, and from one-fourth to one-half of those whose worry about drugs was in the lowest category were, nevertheless, not drug users. Therefore, neither high knowledge nor low worry alone appears to be necessarily predictive of drug use. It is therefore important to explore the possible interaction of these two factors as potential predictors of use.

The interaction between drug use, information, and worry was assessed at follow-up using the Kullbach information ratio. Follow-up data were used on the assumption that they offered the best opportunity to assess residual effects of the program. Table 4 demonstrates that (*a*) use, worry, and knowledge are not independent and, there-

TABLE 2

LEVEL OF INFORMATION ABOUT DRUGS AMONG SUBJECTS WHO DO AND DO NOT USE ALCOHOL, MARIJUANA, AND LSD

Level of information	Alcohol				Marijuana				LSD			
	Do not use		Do use		Do not use		Do use		Do not use		Do use	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Low ($\bar{X}$ = 28.96, SD = 7.74)	94	39.7	185	33.5	165	26.4	33	20.9	259	35.4	19	35.2
Medium ($\bar{X}$ = 41.86, SD = 2.59)	77	32.5	185	33.5	180	28.8	36	22.8	246	33.7	16	29.6
High ($\bar{X}$ = 51.66, SD = 4.55)	66	27.8	183	33.1	281	49.9	89	56.3	226	30.9	19	35.2
Chi-square (df = 2)	3.3048				6.6263*				.5383			

* $p < .05$.

fore, significantly covary; (*b*) only among LSD users do worry and knowledge covary; (*c*) the 2 × 2 tables formed by dividing users of alcohol, marijuana, and LSD according to whether they are high or low in worry and knowledge all yield significant differences; (*d*) level of knowledge alone is insufficient for predicting use; and (*e*) use can be predicted from an interaction between low worry and high knowledge, with high worry–high knowledge *not* being predictive of use.

DISCUSSION

Before discussing the results of this study, seven caveats should be noted. First, it should be borne in mind that the data for this research were drawn from anonymous questionnaires and that use data were wholly self-reports. Anonymous questionnaires were selected for two reasons. First, it was believed that they were the only means through which students would honestly answer questions about their illegal use. Second, the anonymity of responses precluded their being subpoenaed in any court proceedings. Self-report questionnaires were used because they were far more practical than peer or police reports or urine, blood, and other physical analyses (which would pose a serious civil rights dilemma), which would have been alternative means of determining use. It is, of

TABLE 3

LEVEL OF WORRY ABOUT DRUGS AMONG SUBJECTS WHO DO AND DO NOT USE ALCOHOL, MARIJUANA AND LSD

Level of worry	Alcohol				Marijuana				LSD			
	Do not use		Do use		Do not use		Do use		Do not use		Do use	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Low ($\bar{X}$ = 10.05, SD = 7.74)	64	26.0	203	34.5	179	27.0	98	57.5	234	30.1	33	56.9
Medium ($\bar{X}$ = 15.81, SD = 1.49)	63	25.6	203	34.5	238	35.8	63	37.1	241	31.0	23	39.7
High ($\bar{X}$ = 18.55, SD = 2.61)	119	48.4	183	31.1	247	37.2	9	5.3	302	38.9	2	3.4
Chi-square (df = 2)	22.512*				83.319*				31.977*			

* $p < .001$.

TABLE 4

INFORMATION RATIO ANALYSIS OF DRUG USE AS A FUNCTION OF WORRY AND KNOWLEDGE ABOUT DRUGS AT FOLLOW-UP

Source	*df*	Alcohol (n = 322)	Marijuana (n = 324)	LSD (n = 322)
Use: worry: knowledge	4	19.0084*	51.5921*	44.8164*
Worry: knowledge	1	1.9875	1.3417	23.1437*
Use: worry, knowledge	3	17.0209*	50.2504*	21.6727*
Use: knowledge	1	.2002	1.6498	1.7777
Use: worry/knowledge	2	16.8207*	48.6006*	19.8950*
Use: worry/low knowledge	1	2.6529	2.8514	.4631
Use: worry/high knowledge	1	14.1678*	45.7492*	19.4319*

* $p < .001$.

course, possible that anonymous self-report answers, cloaked by their nonaccountability, could be capricious in addition to being subject to biased recall. Despite these risks, practicality demanded the use of both expedients, but their threats to the validity of the data should not be forgotten.

The second qualification worth noting is that the results of this investigation may not generalize beyond the work of the two teachers who delivered a specific curriculum to students in two grade levels of a dominantly upper-middle-class, academic community. Without replication by other educators, using different materials in a broad array of settings, the results reported here must be regarded as little more than suggestive.

Third, these findings should be considered in light of possible response biases, which might have accounted for some of the observed differences. All data were collected from experimental and control subjects by the two teachers. It is quite possible that subjects in the experimental group who had 10 contact hours with the drug educators by the time of posttest and follow-up might have been more honest in their responses than were the controls with whom the teachers were relatively unfamiliar. This could account for the differences that appeared following the initial testing. However, at pretest, experimental subjects showed higher drug use, knowledge, or scores on attitudes conducive to drug use. This would suggest that a response bias was established merely by virtue of assignment to a drug education class. lending some credence to the possibility that merely becoming the target of drug education could intensify the possibility that drugs might be used. In future studies, it is important to avoid the possible confounding of results, which might have occurred here by arranging for the collection of evaluation data by researchers who were not directly associated with the instructional program. Furthermore, the fact that subjects were exposed to the instrument as many as three times might also have introduced a bias in the results, which could have led both experimental and control subjects to admit to higher levels of use and sale. While the results that were obtained closely correspond to the survey data taken from comparable populations in the same community one year earlier, inclusion of a posttest-only condition could have obviated this as a possible problem.

Fourth, the conclusions of this research do not assess the impact of the program upon the patterns of use by subjects who have taken amphetamines, barbiturates, or narcotics. It is plausible that while participation in the program might have hastened some subjects' experimentation with alcohol, marijuana, or LSD, other participants in the same program might have diminished their use of the first three classes of drugs. Data to answer this question are unavailable in the present study because of the insufficient number of users of these drugs found in the junior high school sample. The research was undertaken in junior high schools because of a belief that it would be more beneficial to attempt to interrupt drug use earlier rather than later in the careers of

users, and because the school system in which the study was conducted regularly requires drug education at this academic level. In the interests of a more comprehensive evaluation, however, future researchers might be advised to select subject populations known to represent a broader spectrum of drug users.

Fifth, it is important to bear in mind that the subjects in this study were relatively slight users of alcohol, marijuana, and LSD. Inspection of Table 1 reveals that of the three drugs in question, average use levels both before and after the educational program were at one to two times per month for alcohol and one to two times per year for marijuana and LSD. Compared with reports of drug use in other localities, these are modest levels. Changes in use were likewise of a low average level. Therefore the findings of this research may also not generalize to populations that include comparatively larger numbers of committed users.

Sixth, the follow-up period in this research was a brief four months. It is possible that a longer follow-up period might have revealed more positive changes. However, inspection of Figures 1 and 2 suggests that sale and use of drugs by experimental subjects were still rising at the time of follow-up, creating the possibility that a longer follow-up period might show even more negative changes.

Finally, the results of this study cannot be generalized to all forms of drug education. The present effort was exclusively devoted to imparting drug-related information. Programs that are oriented to promoting value change, to enhancing social or academic functioning, or to other indirect approaches to drug use may have entirely different effects than do those encountered in the present investigation.

With these qualifications in mind, three principal conclusions were drawn from this research. First, it was concluded that drug education may not only impede the use of drugs, it may actually exacerbate drug use. As a means of cross-checking the data reported earlier, subjects were asked at pretest, posttest, and follow-up whether their use of drugs was greater than, equal to, or less than it had been three months earlier. While 5.4% of the experimental subjects and 3.7% of the control subjects reported escalating drug use at pretest, these percentages rose to 13.0% and 4.1%, respectively, at follow-up. Subjects' self-assessment thus appeared to corroborate judgments drawn from other descriptive measures. Threats to the validity of this conclusion based upon the historical fact of passage of the local marijuana ordinance were assumed to be controlled by the use of random assignment of subjects within the experimental design, and the contamination of experimenter–subject interaction was hopefully lessened after the interval of the follow-up period.

Related to this first conclusion is the ap-

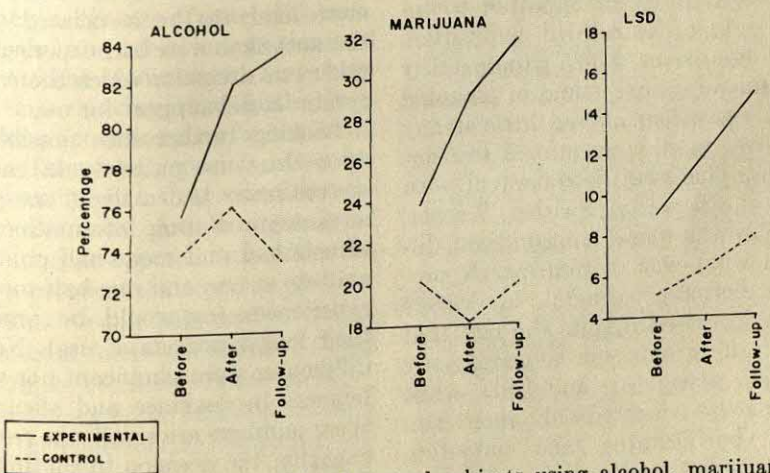


FIGURE 2. Percentages of experimental and control subjects using alcohol, marijuana, and LSD before and after drug education and at follow-up.

parent fact that participation in a drug education program might lead to a dispersion of effects in other behavioral and attitudinal areas. For example, inasmuch as it was found that sellers of drugs had significantly greater ($t = 4.19, df = 814, p < .001$) knowledge than had nonsellers ($\bar{X} = 42.35$, $SD = 14.83, n = 67; \bar{X} = 36.98, SD = 9.59, n = 749$, respectively) and that they also had significantly less ($t = 12.49$, $df = 923$, $p < .001$) worry about drugs than had nonsellers ($\bar{X} = 12.49$, $SD = 10.68$, $n = 72$; $\bar{X} = 15.829$, $SD = 11.37$, $n = 853$, respectively), changes mediated by drug education might well eventuate in effects far beyond the expectations of the drug educator. In light of this possibility, it is wise for drug educators to carefully consider the wisdom of undertaking any program that might disturb decisions made at the natural choice points of the paths leading away from or toward drug use.

The second important conclusion reached by this study was that classroom instruction can lead to a significant increase in the level of students' knowledge about drugs. It has been found that students undergoing drug education learned significantly more than did controls who did receive some drug education through other less concentrated programs in the two schools, through drug facts frequently discussed in mass media, on the street, or through any literature that they may have sought on their own initiative. More important, however, is the finding that neither the format nor the content of the instructional program appeared to influence the rate of knowledge acquisition. Student-led discussions, which paradoxically require considerably more time in planning and skill in execution, offered little if any advantage over more conventional teacher-led programs. This result is consistent with the findings of others (e.g., Swisher, Warner, & Herr, 1972) who found no significant differences following even more disparate programs. Furthermore, general knowledge about drugs increased despite the fact that one-third of the group was educated only about alcohol, marijuana, and LSD, while another third was educated only about amphetamines, barbiturates, and narcotics, with the third subsample receiving education about all six classes of drugs. Unfortunately, the drug information measure used in this research assesses general information and does not lend itself to a dichotomization of content paralleling the divisions in instructional content. Future researchers might be advised to plan for such an evaluation by matching units of instructional content with sectors of the evaluation instruments.

The third major conclusion of this research was that the association between drug education and drug use may not be a simple one. Drug education was observed to lead to greater knowledge about drugs, but increased knowledge alone was not predictive of intensified use. One of the important mediating variables in the chain leading from information to use appears to be the level of worry about bad drug effects. But high knowledge does not necessarily lead to low worry and low worry does not necessarily lead to high drug use. Therefore other factors may also be influential. In the present sample, data from subjects who indicated that they had experienced one or more bad drug effects were analyzed to isolate factors associated with decisions to continue or discontinue drug use following these experiences. Among users of alcohol, marijuana, and LSD, 18.7% ($n = 109$), 23% ($n = 41$), and 47.9% ($n = 35$), respectively, reported having had bad experiences. Of these, 23.9% of those who had bad experiences with alcohol discontinued use as did 17.1% of marijuana users and 37.1% of LSD users. Therefore bad experiences with LSD were clearly more likely to be associated with discontinuance than were bad experiences with the other two drugs for which there is somewhat greater social support for use.

Looking further into possible influences upon drug use patterns, in each instance, discontinuers had a slight one point or less advantage on drug information, while continuers had under one-half point less worry and up to one and one-half more bad drug experiences (as would be predicted from their longer periods of use). None of these differences were significant nor were the differences in deviance and alienation. While these numbers are small, the results do indicate that the decision to continue or discontinue drug use, and presumably to refrain from or initiate use, is controlled for this

population by factors other than those identified in the present research.

There are a great many places to look for the conditions controlling drug use. Some are doubtless related to situational factors such as drug availability and patterns of peer use, the richness of alternatives to drug use mediated by parents and teachers, and the type of health and risk-taking education which the student may have undergone. Another influence may be the drug experience itself. For many, patterns of the use of most drugs tend to be mild and of short duration (National Commission on Marijuana and Drug Abuse, 1972, Appendix 1, pp. 257–258; Stuart & Schuman, 1972, pp. 90–117). When Stuart and Schuman (pp. 136–138) asked subjects their reasons for discontinuing use, "No further need for this experience" was chosen significantly more often than such alternatives as fear of legal, physical, or psychological consequences. Finally, factors relating to brain chemistry and other physiological characteristics may influence the decision to search for or continue to use selected drugs.

Drug education alone should not be expected to reverse the impact of these multiple determinants of drug use, despite the zeal with which many drug educators have responded to community requests for action. The attempt to educate the citizenry rather than to change the controlling conditions is a familiar response to social crises in contemporary society. Regrettably, educational efforts relating to the use of auto seat belts, smoking, alcoholism, racial integration and preservation of the environment have been all-too-slow in achieving their goals. It is likely that drug education may be similar in its results. Therefore, special attention should be given now to drug education research that compares the effectiveness of drug information as opposed to health education or to a program that contrasts educational programs on one side and efforts to correct situational forces conducive to drug use—for example, sterile academic environments, unreinforcing home experiences, and/or a paucity of constructive behavioral alternatives in the community—on the other. Furthermore, the outcome measures for these comparative studies should be fully triangulated to include an assessment of drug use and factors believed to be related to drug use in addition to any consumer evaluation of the program that might be deemed necessary. Only in this way can the critical questions about the efficacy of drug education be answered. And as these questions are answered, so, too, will be questions about "the effectiveness of education itself as a tool for countering social disorganization and decay [Richards & Langer, 1972, p. 48]."

REFERENCES

California Department of Education. *A study of more effective education relative to narcotics, other harmful drugs and hallucinogenic substances.* Sacramento: California Department of Education, 1970.

California State Assembly. *Bill 1730,* Chapter 1437, Section 1, 1967.

Ford Foundation. *Drug abuse survey project.* New York: Praeger, 1972.

King, R. Anonymous versus identifiable questionnaires in drug use surveys. *American Psychologist,* 1970, **25,** 982–985.

Kullbach, S. *Information theory and status.* New York: Dover, 1968.

Kullbach, S., Kupperman, M., & Ku, H. H. Tests for contingency tables and Markov chains. *Technometrics,* 1962, **4,** 573–608.

Michigan State School Code, 1955, 15.3364, Section 364.

National Commission on Marijuana and Drug Abuse. *Marijuana: A signal of misunderstanding.* Washington, D.C.: U.S. Government Printing Office, 1972.

Richards, L. Psychological sophistication in current drug abuse education. In J. R. Wittenborn (Ed.), *Communication and drug abuse.* Springfield, Ill.: Charles C Thomas, 1970.

Richards, L. G., & Langer, J. H. *Drug-taking in youth.* Washington, D.C.: U.S. Government Printing Office, 1972.

Stachnik, T. J. The case against criminal penalties for illicit drug use. *American Psychologist,* 1972, **27,** 637–642.

Stuart, R. B., & Schuman, M. C. *Tripping and toking in mid-America: A survey of teenage drug abuse in four Michigan communities.* Lansing, Mich.: Office of Drug Abuse and Alcoholism, 1972.

Swisher, J., Crawford, J., Goldstein, R., & Yura, M. Drug education: Pushing or preventing? *Peabody Journal of Education,* 1971, **55,** 68–75.

Swisher, J. D., Warner, R. W., & Herr, E. I. Experimental comparison of four approaches to drug abuse prevention among ninth and eleventh graders. *Journal of Counseling Psychology,* 1972, **19,** 328–332.

(Received January 29, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 202-211

LEARNING FROM PROSE TEXT:

EFFECTS OF READABILITY LEVEL, INSERTED QUESTION DIFFICULTY, AND INDIVIDUAL DIFFERENCES[1]

JACK H. HILLER[2]

Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, California

Immediate and two-week retention were studied as a function of three levels of text readability and two levels of inserted postquestion difficulty. The inserted postquestion treatment was modified to permit review of text after question answering. A traditional control group was required to read without marking lesson pages; a second control group was permitted complete freedom. Both inserted postquestion treatments produced significantly inferior acquisition of content incidental to the inserted postquestions for two lessons having below-average readability. For an average readability lesson, only the difficult inserted postquestions produced lower acquisition. Treatment differences diminished to nonsignificant levels on two-week retention. Learning was correlated with anxiety and self-confidence in the two lower-than-average readability lessons but not in the average one.

Numerous studies have investigated the effects of adjunct questions on learning from prose text. A set of widely held conclusions is that inserted postquestions may raise the learning of (*a*) specific details asked about by the questions (Rothkopf, 1966), (*b*) details categorically related to the questions (Rothkopf & Bisbicos, 1967), and (*c*) information not directly related to the questions (Rothkopf, 1966). When a lesson contains more information than a student can learn during a given study period, specific questions surely enhance question-related learning by directing and focusing attention and by stimulating rehearsal. The intriguing experimental result has been that incidental learning is not significantly reduced as a consequence of inserted postquestion treatments; therefore, incidental learning alone was investigated in this study.

Possible bases for the effects of adjunct questions have been suggested by Frase (1970): "Questions are motivational stimuli. They have arousal and associative outcomes [p. 346]." Arousal may be related to drive level, and theorists have long held that the relationship between learning performance and drive is curvilinear and that it is a function of task complexity (Yerkes & Dodson, 1908). Hypothetically, arousal level varies with perceived question difficulty, and task complexity varies with text readability. To enable a test of the hypothesis that learning is a complex interactive function of such text and question characteristics, two sets of inserted postquestions having significantly different difficulty levels (produced by manipulation of distractors for multiple-choice items) were varied factorially against a lesson written

[1] An earlier version of this article was presented at the meeting of the American Psychological Association, Honolulu, September 1972. The data were collected while the author was at Southern Illinois University, Carbondale. The author acknowledges his indebtedness to E. Z. Rothkopf and L. T. Frase for their 1970 American Educational Research Association presession which stimulated this research, and he thanks George Klare whose research on readability provided the model. The author also expresses his appreciation to Ernst Rothkopf, John B. Carroll, Edys Quellmalz, and Reed Williams for valuable comments on the earlier manuscript.

[2] Requests for reprints should be sent to Jack H. Hiller, Army Institute for Research in Behavioral Science, ARI/CDEC, Fort Ord, California 93941.

at three levels of readability (produced by the manipulation of lexicon and syntax). A variety of aptitude measures, described below, were gathered and correlated with test results to aid in the interpretation of treatment effects.

A second question of this study concerned the practical utility of conclusions based on previous research dealing with the effects of adjunct questions. Standard control groups have been instructed to read passively. However, observation of the study habits of college students indicated that passive reading is atypical; students may instead be observed to underline, highlight, write notes, and outline (normative questionnaire data were collected to estimate frequencies). Students in one control group of this experiment were, therefore, directed to study according to their idiosyncratic habits. A second control group was given the traditional direction to read only.

The external validity of research on prose study techniques also requires testing of the experimental techniques as they might actually be utilized by students. However, in contemporary research on the use of adjunct questions, the dominant technique that has been employed for theory development is one that students would surely disregard if a course grade or graduation were contingent on achievement. Specifically, subjects have been prevented from looking back at the lesson once they have seen the postquestion; they have also been prevented from reviewing the lesson prior to administration of the criterial test. A second problem confronting attempts to generalize from the existing literature to applied practice concerns the fact that experimenters have neither explained the purpose of the inserted questions to their subjects, suggested how the questions might be used to improve study effectiveness, nor explicitly described the relationship between the inserted questions and the retention test. These problems were dealt with in the present study.

Experiments designed to improve or create new instructional techniques must also employ study materials similar in character to those actually in use, or external validity is damaged. Furthermore, it may be speculated that improved instructional techniques are most needed for course materials that students find difficult and uninteresting. Given these considerations, an effort was made to select a lesson that students would find boring and hard to study.

Method

Lessons

The introductory section of a college-level mathematics text was selected. It contained a description of the abstract, deductive character of modern mathematics, a sketch of its history, and the authors' philosophical perspectives. The reading difficulty level was obviously high because of difficult vocabulary and syntax. Of a 68-sentence total, 29 sentences contained 25 or more words. The text contained 1,687 words, and the average number of words for all independent clause units (or simple sentences) was 22.0.

Two versions of the original text were prepared by breaking long sentences into shorter ones and by replacing uncommon terms with more familiar synonyms. In addition, a concrete example was introduced to clarify an abstract discussion. A moderately revised version contained 1,874 words and its independent clause units averaged 18.4 words. A highly revised version contained 1,961 words and averaged 14.9 words per clause. Hereafter, the original lesson is referenced as having low readability (that is, as being very difficult), the moderate revision as having moderately difficult readability (moderately difficult), and the major revision as having average readability (average difficulty). Subjective scaling data and results for comprehension testing which support these descriptions are reported.

The lessons were collated into 11-page booklets consisting of a cover page containing general experimental directions, a second page explaining the particular study technique, and a 9-page text, with the exception that booklets containing inserted question treatments had 9 additional pages for the questions and a blank page placed between each text page and its associated question.

Treatments

Four treatments were employed: relatively easy inserted postquestions, relatively difficult inserted postquestions, passive reading, and idiosyncratic study.

Inserted postquestions. The student was instructed to study each lesson page until confident that he could answer a question on its contents. It was explained by the directions that the "questions are designed to serve you as check points" and that answers were not given for two reasons:

(1) the answers are in the lesson material just read, and you may look back if you have any doubts concerning the correct answer, (2) previous research has shown that most students simply look for the answers provided rather than studying. If you feel reasonably confident about the answer to each question, then you have a sign that your progress is adequate. On the other hand, if you are not sure and have to keep looking back, then you have a sign to slow up and study harder.

Although students were permitted to check their answers, they were required to respond on the data sheets before turning back and were barred from erasing to preserve this data for analysis.

It seemed necessary to divide the text into relatively small sections so that limitations imposed by immediate memory capacity would not make the study technique unmanageable. With sufficiently small amounts to be learned for each postquestion, it was hoped that students would be encouraged to rehearse in preparation for the unknown question and that they would thereby improve retention. Based on this consideration, the original text was divided into nine sections with a median length of 199 words and with the division points placed carefully to avoid breaking continuity in thought (divisions were at paragraph boundaries in eight instances). The section length of about 200 words was also chosen because it corresponded to the interval that Frase (1968) had found optimal for raising retention on a combined test of incidental and repeated questions. Both revised lessons were divided at the same points as the original lesson.

A single multiple-choice question stem was written in the knowledge or low-level comprehension domains of the Bloom (1956) taxonomy for each of the nine sections. The questions may also be described by Anderson's (1972) categories as containing one verbatim item, four transformed verbatim items, and four transformed paraphrase items. These questions were specifically, rather than randomly, selected to prevent the occurrence of any spurious patterns. Two sets of adjunct questions were formed by manipulating the semantic relatedness of the distractors (three). One set of adjunct questions was organized to contain all of the easy items and the other set was organized to contain all of the hard ones. The adequacy of this manipulation was directly tested by examining the item difficulty indices calculated from data provided when the students recorded their answers on data sheets during the experiment.

Passive reading. Students worked according to the following instruction: "The method you are to use is the common study technique of reading. You may study in any manner that you please; only do not use your pen or pencil as a study aid."

Idiosyncratic study. The direction was to "study just as you do typically. If you like to underline, or pen notes in the margins, or just read straight, then do so now."

Questionnaire

Immediately after studying the experimental lessons, but prior to taking the test, the students were given either a blank sheet on which to write opinions or a questionnaire which asked about (*a*) interest in the lesson topic, (*b*) readability level of the text, and (*c*) merit of the experimental study method. (The questionnaire was given before testing to avoid possible artifactual effects; for example, those who studied inserted questions might have reacted negatively when finding them missing from the test.) Ratings for these questions were secured through use of 5-point rating scales. The students were also questioned about their study habits.

Tests

Acquisition and retention. A multiple-choice test composed of 28 items, with four choices per item, was given immediately after the questionnaires were collected. Two weeks later, the same test was administered without prior warning to students having studied the very difficult lesson. Students having studied the moderately difficult and the average lessons were given 10 minutes to restudy their lesson booklets prior to retaking the test; this procedure was designed to explore the possibilty that note writing and underlining of lesson material facilitates review activities and to improve external validity, on the assumption that most students study or review before taking a test. Twenty minutes were provided for both immediate and retention testing.

Transfer study. The possibility that the inserted questions could provide direct positive transfer to the criterial test, which contained items similar in type to the adjuncts but did not contain any of them, was investigated. A group of 21 students first took the test without any preparation. After this pretest, they were given handouts containing the adjunct question stems and answers and were then retested with the handouts as an available aid.

Test fidelity. The criterial test had been constructed directly from the original lesson. To check the possibility that test information was not adequately preserved or represented in the two revisions, a separate experiment was conducted. Forty students in two course sections of undergraduate educational psychology were randomly assigned to one of three groups, each of which studied one of the three lessons for 25 minutes. These students were led to believe that they would be given a closed-book test to insure that they had acquired a thorough acquaintance with the lesson contents. However, an immediate open-book test (20 minutes completion time) was given instead to minimize the possible effects of writing style on retention and to reflect more accurately the available text information. By accepting the assumption that text information was preserved, mean test scores can also be used to index the

changes in lesson comprehensibility brought about by the text revisions.

Participants

Students of the undergraduate educational psychology course winter and spring quarters, 1971–72, at Southern Illinois University, Carbondale, participated to fulfill a research requirement. Data were collected from approximately 700 students, but certain losses occurred. It was discovered after data collection that 15 booklets for the very difficult lesson, hard-question treatment were improperly collated. A number of students for whom immediate retention data had been collected were absent during the second-week retention-testing periods. Attitude data were missing for a small percentage of students, but American College Testing Program, English subscale data were missing for about one third of the sample. To improve the power of significance tests, the largest sample available for each analysis has been used. As a consequence, means, standard deviations, and sample sizes are reported for the several analyses presented in Tables 1, 2, and 3 of the Results section. It may be noted that the sample sizes of the idiosyncratic treatment for the very difficult and moderately difficult lessons are relatively low; less data had been collected here since the booklets for the other treatments could be reused. It should also be observed that more data were collected for the very difficult than the moderately difficult lesson and for the moderately difficult than the average lesson, in accordance with the major interests of this experiment.

Aptitude Measures

Verbal ability scores—American College Testing Program, English subscale—were obtained from administrative records for the majority of students. A variety of attitude measures were obtained by the use of a battery administered during the first class meeting in both quarters. The specific scales were as follows: Alpert-Haber (1960) Achievement Test Anxiety, Facilitating and Debilitating subscales; Internal–External Locus of Control (Rotter, 1966); Dogmatism (Rokeach, 1960); Social Desirability (Crowne & Marlow, 1964); and intellectual self-confidence, a scale developed by the investigator which is described below.

The intellectual self-confidence scale is based on a construct defined as follows. Phenomenologically, the *belief* that one has the capacity to succeed at tasks demanding intellectual effort is the central fact of intellectual self-confidence. Theoretically, the strength and scope of this belief is a function of the individual's reinforcement history. A history of actual, or perceived, success would shape a positive conviction and failure a negative self-regard. In addition, if it is assumed that success constitutes positive reinforcement, then cues associated with intellectual activity themselves acquire reinforcing properties (i.e., become positive secondary reinforcers). Thus, the successful individual develops a liking for intellectual activity. Furthermore, since the successful individual has been reinforced for his own efforts, self-reliance also has been shaped. The construct definition for intellectual self-confidence, therefore, incorporates reference to three behavioral tendencies: (*a*) expectation of success, (*b*) attraction to intellectual tasks, and (*c*) self-reliance. The 33 items of the intellectual self-confidence scale have been written to provide a nonreactive measure of these three component tendencies. Scale reliability is approximately .75, Cronbach α. Validation information is reported in Kirby and Hiller (1973) and Cronbach and Snow (1974).

Procedure

Treatments were assigned randomly to students within intact classrooms (28 in the main experiment) with the restriction that each class was to have only one of the three lessons. The experiment was conducted during regular class hours. The goal of this experiment, "to develop information on the effectiveness of different study techniques," was described to students on the cover of the lesson booklets. The different treatments were not, however, described to the classes. To foster academic motivation, the instructor informed students that he would be apprised of any students scoring too low (below chance level) for credit. The final paragraph of the lesson cover stated that there would be 25 minutes of study time and that the lesson contained about 1,800 words: "This means that you will have time to read it twice over—in other words you have enough time to *study*." (Only a negligible number of students were observed to be studying actively when the 25-minute period ended, so that unavailability of time was not a factor. Questionnaire responses further indicated that lack of study time did not differentially affect study behavior among treatment groups.) The students were informed just prior to getting their study directions that the test was designed to "test your retention of facts and your understanding of the ideas presented in the lesson." After reading the second page of their booklets, which contained directions for the study techniques, the students began to study.

Results

Transfer Study

The group tested to determine if the inserted questions provided direct positive transfer to the test averaged 7.5 on the pretest and 7.9 on the posttest ($t = .24$, $df = 20$). Hence, the criterion test measured retention of information not directly related to the inserted questions.

Test Fidelity

The groups given the open-book test to determine if test-related information had been adequately preserved in the two revised lessons performed as follows: very difficult lesson, $M = 15.9$, $SD = 3.5$, $n = 11$; moderately difficult lesson, $M = 17.6$, $SD = 2.5$, $n = 18$; and average lesson, $M = 19.5$, $SD = 2.7$, $n = 11$. The test means were significantly different ($F = 4.63$, $df = 2/37$, $p < .05$). In addition, the difference between results for the moderately difficult and the average lessons was significant ($t = 2.07$, $df = 27$, $p < .05$) under a nondirectional test. Given the open-book nature of the test, results here may be interpreted as providing an objective measure of lesson comprehension, thus verifying the effectiveness of the readability manipulations.

Inserted Question Difficulty

The easy inserted questions achieved an average difficulty level across the three lessons of 69%, while the hard questions averaged 53% ($F = 65.3$, $df = 1/323$, $p < .001$). Neither lesson differences nor the interaction between question treatments and lessons approached significance.

Immediate Retention

Analysis of immediate test results (see Table 1) demonstrated that idiosyncratic study and passive reading were similar to each other but superior to the inserted question treatments. Overall analysis of variance for the experiment determined that treatment effects were highly significant ($F = 11.6$, $df = 3/663$, $p < .001$).[3] Lesson effects also were found to be significant ($F = 6.86$, $df = 2/663$, $p < .01$). The interaction between lessons and study treatments was not significant ($F = 1.08$, $df = 6/657$) for the experiment as a whole. However, inspection of immediate retention means suggests that lessons and treatments did interact for the inserted questions, but that uniform performance for passive reading and idiosyncratic study obscured the interaction analysis. In fact, separate analysis for the inserted question data does indicate a tendency toward interaction ($F = 2.78$, $df = 2/321$, $p < .06$).

In light of the increase in comprehension scores across the lessons, it is interesting to observe that immediate retention increased from the very difficult lesson ($M = 12.5$, $SD = 3.1$) to the moderately difficult lesson ($M = 13.5$, $SD = 3.8$), but that immediate retention for the average lesson ($M = 13.5$, $SD = 3.3$) did not increase further.

Two-Week Retention

Retention data (see Table 1) demonstrated no significant effects for the very difficult lesson analyzed separately or for the revised lessons analyzed together. Since results for the immediate test indicated that the inserted questions lowered the retention of information incidental to the questions and that the two control groups learned about equally well, two-week retention data have also been analyzed separately for the control and experimental groups using the two revised lessons. Results for the inserted question groups did not demonstrate significant differences in retention, but the controls did differ significantly across the moderately difficult and the average lessons; retention was higher for the moderately difficult lesson ($t = 2.63$, $df = 150$, $p < .05$). Parallel to this result, the controls did not show a significant loss of retention in the moderately difficult lesson, while controls using the average lesson did drop.

Aptitude–Retention Correlations

Correlations between aptitude and immediate retention scores are displayed in Table 2, and summary statistics are displayed in Table 3. In the very difficult lesson, the measure of academic confidence (intellectual self-confidence scale) correlated significantly with scores of students having passive reading and inserted question treatments and was, in fact, most highly correlated with performance by the hard-question group. Furthermore, in the moderately difficult lesson, the intellectual self-confidence scale correlated only with scores for the hard-question group, while in the average

[3] The analysis of variance was performed by the multiple regression technique and program described in Kelly, Beggs, and McNeil (1969).

TABLE 1
TEST RESULTS

Treatment	Lesson								
	Very difficult			Moderately difficult			Average		
	M	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Immediate retention									
Passive reading	13.1	3.3	86	14.4	3.9	68	13.7	3.2	35
Idiosyncratic	13.3	2.1	59	14.4	3.7	41	14.2	3.3	53
Easy questions	12.1	3.1	81	12.2	3.3	60	13.4	3.3	34
Hard questions	11.4	3.1	64	13.1	3.9	55	12.2	3.0	33
Two-week retention									
Passive reading	11.1	3.1	71	13.7	4.9	54	11.6	4.1	27
Idiosyncratic	11.1	2.4	47	13.7	4.0	33	11.9	3.2	38
Easy questions	10.6	3.3	69	11.9	3.8	47	11.4	3.5	24
Hard questions	10.2	2.9	53	12.9	4.3	46	12.2	3.3	25

lesson, it correlated with none. The facilitating test anxiety measure correlated with both control group results and with those for hard questions in the very difficult lesson, but was not correlated for any other treatment conditions. Debilitating anxiety produced correlations that mirrored those for the intellectual self-confidence scale. Verbal ability was generally correlated with performance for the various treatments and showed a tendency in all three lessons to be more predictive of performance for passive reading than for idiosyncratic study (see Footnote *b* in Table 2).

Results for Dogmatism, Social Desirability, and Internal–External Locus of Control scales were not generally significant and, therefore, are not reported here.

Two-week retention–aptitude correlations for the identical students employed in the immediate retention analyses are displayed in Table 2. Here the intellectual self-confidence scale emerged as a significant correlate for the idiosyncratic treatment of the very difficult lesson, while the correlation for hard question dropped from .43 to .26. In the moderately difficult lesson, correlations for passive reading and easy question were raised to significant values, while the correlation for hard question dropped from .41 to .18. Facilitating anxiety maintained significant correlation in the very difficult lesson only for passive reading but rose to significance for the idiosyncratic study treatment of the moderately difficult lesson.

Interrelatedness of the aptitude variables is shown by the following correlations (N = 427): Intellectual Self-Confidence × Debilitating Subscale = −.37; Intellectual Self-Confidence × Facilitating Subscale = .23; Intellectual Self-Confidence × American College Test = .31; Debilitating Subscale × Facilitating Subscale = −.39; Debilitating Subscale × American College Test = −.27; Facilitating Subscale × American College Test = .20.

Data were graphed and examined for disordinal aptitude–treatment interactions. Nonsignificant but consistent patterns suggested that low-aptitude students performed best under idiosyncratic directions, while high-aptitude students performed best under passive reading in the very difficult lesson (the general lack of significant correlations in the two revised lessons led to a nonsignificant F test of interaction for all lessons, treatments, and aptitude variables combined in one analysis).

Questionnaire

Readability. Subjective ratings for lesson readability demonstrated that the text manipulations were effective (F = 20.6,

TABLE 2

IMMEDIATE- AND TWO-WEEK-RETENTION–APTITUDE CORRELATIONS

Lesson/treatment	n[a]	Correlations								
		Immediate retention-aptitude					2-week retention-aptitude			
		ISCS	FA	DA	ACT-E[b]	n	ISCS	FA	DA	ACT-E
Very difficult										
Passive reading	63	.30*	.32*	−.25*	.28	43	.32*	.26*	−.18	.28
Idiosyncratic	44	−.02	.32*	−.10	.45*	38	.38*	.06	−.11	.25
Easy questions	61	.31*	−.04	−.21	.66*	45	.17	.02	−.13	.42*
Hard questions	51	.43*	.33*	−.43*	.30	37	.26*	.23	−.41*	.55*
Moderately difficult										
Passive reading	51	.21	.10	−.20	.37*	37	.30*	.17	−.05	.53*
Idiosyncratic	30	.13	.22	−.23	.56*	23	.33	.39*	−.26	.56*
Easy questions	38	.12	−.11	.07	.25	27	.37*	.11	.05	.36
Hard questions	44	.41*	.11	−.45*	.29	36	.18	.01	−.28	.33*
Average										
Passive reading	25	.09	.13	.07	.41	21	−.11	.31	−.14	.53*
Idiosyncratic	34	.12	.12	−.25	.22	31	−.06	.14	−.15	.28
Easy questions	22	.26	.06	−.14	.50	15	.00	.38	−.24	.63*
Hard questions	21	.28	.05	−.29	−.12	16	.27	−.02	−.05	.24

Note. Abbreviations: ISCS = intellectual self-confidence scale; FA = Facilitating subscale and DA = Debilitating subscale of the Alpert-Haber Achievement Test Anxiety Scale; ACT-E = American College Testing Program, English subscale.

[a] Sample sizes refer to all statistics except those for ACT-E.

[b] Correlations between ACT-E and immediate retention scores, based on all data available were as follows (sample sizes given in parentheses): very difficult: passive reading = .46* (50), idiosyncratic = .33* (43), easy question = .60* (52), hard question = .35* (41); moderately difficult: passive reading = .43* (44), idiosyncratic = .33 *ns* (30), easy question = .18 *ns* (35), hard question = .29 *ns* (42); and average is the same as in the above table.

* $p < .05$, nondirectional test.

$df = 2/400, p < .001$); treatment and interaction effects were not significant.

Lesson interest. Interest ratings varied significantly across lessons such that the more readable the lesson, the higher the interest ($F = 4.68$, $df = 2/400$, $p < .01$); treatment and interaction effects were not significant.

Method liking. The students indicated strong differences in their liking for the different study techniques ($F = 12.5$, $df = 3/400, p < .001$); lessons and interaction did not show significant effects. Idiosyncratic study was rated most favorably, followed by easy question, hard question, and passive reading, in that order.

Normative study habits. Approximately 47% of the students reported that they usually write notes and/or underline when studying, but only 17% of those in the idiosyncratic conditions wrote any notes, and most of these stopped writing notes by the third page; 84% of the students reported that they sometimes or usually underline when studying, and 73% of the students in the idiosyncratic condition did underline.

DISCUSSION

The inserted postquestion procedure, as modified to permit text review, failed to enhance retention of content incidentally related to the inserted questions. With the exception of relatively easy questions inserted in the average readability lesson, the postquestion procedure actually tended to depress immediate retention. A possible explanation may be found in the frequency of question pacing, which was roughly one question every 200 words. Frase (1968) found that postquestions inserted in a 2,000-word text every 10 sentences tended to lower incidental learning, but this questioning rate also produced the highest retention scores on a combined test of incidental and inserted question learning and thus appeared to provide an appropriate rate for this experiment.

TABLE 3
MEANS AND STANDARD DEVIATIONS FOR APTITUDE MEASURES

Lesson/treatment	n^a	ISCS		FA		DA		ACT-E		
		M	*SD*	*M*	*SD*	*M*	*SD*	*n*	*M*	*SD*
Very difficult										
Passive reading	63	112.1	13.3	25.0	5.7	27.2	7.2	43	20.8	3.5
Idiosyncratic	44	112.8	14.4	25.6	5.0	26.9	5.6	38	22.0	4.1
Easy questions	61	110.9	11.5	24.4	5.0	28.6	5.9	45	21.3	4.9
Hard questions	51	113.6	12.0	24.2	5.1	27.0	5.6	37	20.4	4.1
Moderately difficult										
Passive reading	51	111.8	14.3	24.4	6.0	27.1	7.5	37	20.1	4.5
Idiosyncratic	30	117.5	10.7	24.5	4.6	27.7	6.0	23	22.8	4.5
Easy questions	38	115.1	15.7	23.2	6.0	28.3	6.6	27	20.8	3.7
Hard questions	44	113.6	15.5	25.9	6.5	27.6	7.2	36	21.0	3.6
Average										
Passive reading	25	107.5	11.7	23.4	4.8	27.9	4.9	21	19.0	3.9
Idiosyncratic	34	112.4	12.2	23.7	5.4	26.8	5.3	31	20.5	3.6
Easy questions	22	111.3	14.3	23.9	5.4	28.2	4.3	15	21.5	3.1
Hard questions	21	111.5	10.6	23.0	4.3	30.0	4.3	16	20.8	3.1

Note. Abbreviations: ISCS = intellectual self-confidence scale; FA = Facilitating subscale and DA = Debilitating subscale of the Alpert-Haber Achievement Test Anxiety Scale; ACT-E = American College Testing Program, English subscale.

[a] Sample sizes apply to all statistics except those for ACT-E.

Frase attributed the incidental learning depression to disruption in the text's continuity. However, in the context of the present experiment, this interpretation must be regarded as equivocal, since the text was divided so as to preserve integrity of communication, and students could review to overcome any difficulties. The pattern of results does suggest two factors that may have contributed to the depression, and these are discussed below.

Immediate retention achieved by the hard question group studying the moderately difficult lesson and by both inserted question groups studying the very difficult lesson was significantly correlated with confidence. An experiment by Means and Means (1971) suggested how confidence (generalized expectation of academic success) may have affected performance. They found that students with low grade point averages (and presumably low success expectation) performed worse in a course as the result of their being led to expect low achievement; in contrast, students with high grade point averages performed above control levels after being informed that they had low aptitude for the course. For the very difficult and moderately difficult lessons, the questions may have achieved effects comparable to the negative manipulation by Means and Means, since performance on the inserted questions was rather poor (on the average, 53% correct for hard questions and 69% for easy ones). Additionally, the significant negative correlation of debilitating test anxiety with immediate retention for the hard-question groups studying the very difficult and the moderately difficult lessons implies that distracting emotional behaviors were evoked (see Wine, 1971).

Examination of results for the average-readability lesson showed that neither achievement test anxiety nor academic self-confidence predicted immediate retention. Since insertion of questions occurred at precisely the same points for both hard- and easy-question treatments and the questions were of the same cognitive type, the fact that hard questions alone produced performance significantly inferior to the idiosyncratic control (Newman-Keuls test, $p < .05$) suggests that question difficulty was the primary factor that contributed to this deficit. The mechanism underlying the deficit might be that students who experienced difficulty with the questions concentrated on question-related material dur-

ing review. In the moderately difficult lesson, immediate retention scores for the easy-question group were not correlated with confidence or anxiety; hence, the significantly inferior acquisition performance by the easy-question group ($p < .05$, compared against the idiosyncratic control by a Newman-Keuls test) may also be attributed to review activities that concentrated attention on question-related information. (Reversal of performance by the two inserted question groups in the moderately difficult lesson is curious, although the difference between these two groups is not significant at $p < .05$.)

The prediction that idiosyncratic study would prove superior to passive reading was not verified. Interfering with normal study habits could be expected to upset some students, and passive reading did produce the lowest treatment evaluation ratings; but the test-anxiety–acquisition and confidence–acquisition correlations only tended to be stronger for passive reading than idiosyncratic study. A possible explanation for the lack of significant effects is derived from an analysis of the usefulness of note taking or outlining in relationship to the purpose of studying and to the contents and organization of lesson materials.

Frase (1969) found that students spontaneously wrote notes when studying a text to perform a task which placed a large strain on immediate memory; however, it was the case here that only about 17% of the students took notes (as compared to questionnaire data which indicated that 88% usually or sometimes write notes), and it was observed from the lesson booklets that note writing typically extinguished by the third page. In addition to the fact that little note writing was done, another point is that Schultz and Di Vesta (1972) found that taking notes was initially advantageous when the text presented an unusual structure (information organized according to attributes rather than names). Both on empirical and logical grounds, it would seem that control subjects should be permitted to study according to individual preference or habit.

One potentially interesting result was that two-week control group retention for the moderately difficult lesson was superior to control retention for the average lesson, despite the fact that comprehension scores and readability ratings for the average lesson were higher. It may be inferred that the moderately difficult readability of the moderately difficult lesson successfully stimulated careful reading.

It should be recognized that the negative findings obtained here for knowledge-level questions may not generalize to the use of other kinds of questions. Some students complained that the questions were concerned with details and suggested that they would have been more interesting and useful had they been aimed at the "main ideas." The finding of Watts and Anderson (1971) that questions which required the application of principles were more effective than questions which required recognition of previously described examples supports this proposal. It is interesting to note that Watts and Anderson raised the possibility that their application questions were more effective simply because they were measurably harder; however, results for the hard-question groups in this experiment imply that additional processing demands rather than difficulty per se explain their results.

To sum up, results for both immediate and two-week retention caution against inserting low-level cognitive questions into text as a means for promoting the learning of information not specifically cued. Furthermore, there yet remains unanswered a most important practical question: How would students actually use inserted questions when studying for school examinations? If the inserted questions were not well represented in their examinations, we may predict that student attention to them would extinguish. On the other hand, high test relevance would encourage careful attention to the questions (or surrogate instructional objectives) during study. Unfortunately, we might then also predict that students would look ahead to find such valuable test clues when studying and thereby risk the loss of incidental learning found by Rothkopf (1966) and others for prequestion treatments. If this analysis is accurate, instruc-

tors wishing to employ adjunct aids would have to construct study questions that comprehensively represent both instructional objectives and examination contents.

REFERENCES

ALPERT, R., & HABER, N. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, 1960, **61**, 207–215.

ANDERSON, R. How to construct achievement tests to assess comprehension. *Review of Educational Research*, 1972, **42**, 145–170.

BLOOM, B. S. (Ed.) *Taxonomy of educational objectives; the classification of educational goals; handbook 1, cognitive domain*. New York: David McKay, 1956.

CRONBACH, L. J., & SNOW, R. Aptitudes and instructional methods. Unpublished manuscript, Stanford University, 1974.

CROWNE, D., & MARLOW, D. *The approval motive*. New York: Wiley, 1964.

FRASE, L. T. Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology*, 1968, **59**, 244–249.

FRASE, L. T. Cybernetic control of memory while reading connected discourse. *Journal of Educational Psychology*, 1969, **60**, 49–55.

FRASE, L. T. Boundary conditions for mathemagenic behavior. *Review of Educational Research*, 1970, **40**, 337–348.

KELLY, F. J., BEGGS, D. L., & MCNEIL, K. A. *Research design in the behavioral sciences, multiple regression approach*. Carbondale: Southern Illinois University Press, 1969.

KIRBY, E. A., & HILLER, J. H. Comparative validation of a direct and an indirect measure of academic self-confidence. Paper presented at the meeting of the American Educational Research Association, New Orleans, February 1973.

MEANS, R. S., & MEANS, G. S. Achievement as a function of the presence of prior information concerning aptitude. *Journal of Educational Psychology*, 1971, **62**, 185–187.

ROKEACH, M. *The open and closed mind*. New York: Basic Books, 1960.

ROTHKOPF, E. Z. Learning from written materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, 1966, **3**, 241–249.

ROTHKOPF, E. Z., & BISBICOS, E. E. Selective facilitative effects of interspersed questions of learning from written materials. *Journal of Educational Psychology*, 1967, **58**, 56–61.

ROTTER, J. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, **80**(1, Whole No. 609).

SCHULTZ, C., & DI VESTA, F. Effects of passage organization and note taking on the selection of clustering strategies and on recall of textual materials. *Journal of Educational Psychology*, 1972, **63**, 244–252.

WATTS, G., & ANDERSON, R. Effects of three types of inserted questions on learning from prose. *Journal of Educational Psychology*, 1971, **62**, 387–396.

WINE, J. Test anxiety and direction of attention. *Psychological Bulletin*, 1971, **76**, 92–104.

YERKES, R. M., & DODSON, J. D. The relations of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology*, 1908, **18**, 459–482.

(Received February 7, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 212-215

ADJUNCT QUESTIONS AND THE COMPREHENSION OF PROSE BY CHILDREN[1]

INGRID SWENSON AND RAYMOND W. KULHAVY[2]

Arizona State University

One hundred and nine fifth- and sixth-grade children ready twenty 66-word paragraphs describing a fictitious island and its people. A critical question on each paragraph was inserted before or after 1, 5, 10, or 20 paragraphs. The retention measure consisted of items testing both critical and incidental material, in both verbatim and lexical paraphrase form. Learners responded to both immediate and one-week retention tests. Postpresentation of questions facilitated learning, and retention loss was greatest for 1-paragraph learners. Critical items were better recalled, and there was no effect for the verbatim–lexical paraphrase variable.

Research on adjunct questions with adult learners has shown that it is often possible to control what a person learns from reading text. The present study attempts to replicate and extend these findings with a grade school population.

Numerous studies have indicated that postquestions facilitate test performance more than prequestions (Frase, 1967, 1968a, 1968b; Rothkopf & Bisbicos, 1967). Interrogatives placed before passages tend to limit inspection to specific (critical) content and to focus attention on the relatively small number of words needed to answer the question. Conversely, questions that appear after passages reduce specific discriminations and promote the acquisition of nonreferent (incidental) material as well. Although these results have occurred consistently with adult populations, we were unable to find any data gathered on school children. Younger learners may react differently to inspection control devices, possibly because of an attenuated attention span or a lack of strongly formed reading behaviors. Consequently, one purpose of this study was to replicate the effects of pre- and postquestions on critical and incidental recall with grade school subjects.

Several studies have also varied the amount of material the subject reads before encountering a testlike event. What effects result when adjunct questions are paced at various intervals is unclear. Shorter text blocks favor the groups that see questions after study, while longer segments benefit learners who receive interrogatives prior to inspection (Frase, 1967, 1968b, 1968c). In addition, effects of question pacing interact with the type of material. For critical material, medium-length passages yield optimum recall. Interestingly, scores on incidental material improve gradually as more information is presented (Frase, 1967). Reading longer passages may be the optimum procedure when testlike events do not occur during study (Frase, 1967). A second objective of this research is to determine if these parameters are consistent using the same text presentation with children.

According to Anderson (1972), there is evidence that students process and store instructional material in at least two ways. The first strategy, *phonological encoding*, refers to verbatim storage of printed verbal stimuli. Here a student learns by "rote" a string of words which is meaningless to him, but which he recalls intact during testing. The second type of processing is *semantic*

[1] This research was partially supported by Arizona State University Faculty Research Grant 7805-837-15-8 to the second author. The authors wish to thank William Greaves and the personnel of the Osborn School District for their cooperation in the collection of data.

[2] Requests for reprints should be sent to Raymond W. Kulhavy, Department of Educational Psychology, Arizona State University, Tempe, Arizona 85281.

encoding which requires that the learner remember the meaning as well as the physical features of the text. In this case, the student "comprehends" the material and can identify instructional statements correctly when they are presented in a form substantively different from the one initially learned. One means of determining if semantic encoding has taken place is to test recall with lexical paraphrases of the material originally learned. A paraphrase is defined as a parallel statement containing the same semantic content in a different substantive form. For example, "the monarch seemed mad at the instructor" might be transformed to "the king appeared angry with the teacher." These representations differ substantively but were judged identical in meaning by almost 90% of a high school sample.

It seems possible that effects of question placement result from differences in the type of coding produced. When questions are placed before reading, they may provide the learner with a phonological "target" on which to expend his search activities. On the other hand, when questions follow reading, the subject may attempt to store general context and, hence, remember semantic structure rather than specific verbal units (Sachs, 1967). A third purpose of the present study, then, was to determine if pre- and postquestion effects are due to differences in the type of coding they produce.

The final variable investigated was retention. If recommendations from adjunct-question research are to be instructionally useful, it is necessary to assess their effects over time. It may be that placement, pacing, and encoding parameters may change appreciably due to factors effecting tenure in long-term storage.

Method

Design and Subjects

The design was a 2 (Before–After) × 4 (1, 5, 10, or 20 Paragraphs) × 2 (Critical–Incidental Item) × 2 (Verbatim–Paraphrase Form) × 2 (Immediate–Delayed Test) with repeated measures on the critical–incidental item, verbatim–paraphrase form, and immediate–delayed test variables. In addition to the eight experimental groups, two control groups were included. One control (Control 1) read the text without questions, and the other (Control 2) answered the questions without reading the passage.

The subjects were 128 students from two fifth–sixth- and one sixth-grade classrooms. Participation was dependent upon class attendance during the two days that the experiment was conducted. Nineteen subjects were dropped from the analysis for failure to complete all experimental tasks.

Materials

A 1,320-word passage titled "The Island of Ako and Its People" constituted the experimental text. The passage was divided into 20 paragraphs, each exactly 66 words in length. For each paragraph, two unrelated questions were constructed. Next, a lexical paraphrase was generated for each of the original questions, thus equaling four questions per paragraph. Each question and its paraphrase had no substantive words in common other than articles and the terminal response. An example of one set of questions for a paragraph is

The islanders construct clothes from *palm leaves.*

Things the natives wear are made with *palm leaves.*

Since making a publicly observable response increases attention to the material studied (Kulhavy & Parsons, 1972), the subjects were required to fill in the terminal blank for each item in the text.

Both the paragraphs and the pairs of questions were normed for readability and semantic similarity, using 127 fifth- and sixth-grade students from the participating district. The norming subjects were familiar with all the text vocabulary except the words coined for places, animals, and plants on the island. The median similarity rating for one set of question pairs was 4.18, and for the second set it was 4.26 on a 5-point scale, with 5 equaling highest similarity.

One of the four possible questions for each paragraph was chosen as the experimental item (to be seen by the subject while reading). The verbatim question, its paraphrase, and the remaining set of questions were included on the criterion test to measure both the encoding strategy used and the amount of incidental learning. The particular items designated as experimental were separately randomized for each booklet, with the restriction that all questions were chosen an equal number of times across conditions.

In the pacing conditions, the subjects read either 1, 5, 10, or 20 paragraphs in conjunction with the same number of associated experimental questions. In the before groups, the appropriate number of questions were presented before the paragraphs, and in the after groups, the questions were given when the subject had completed reading. Subjects were required to complete the blank at the end of each question encountered. The booklets used in the experiment contained only 1 paragraph or question on each page.

The criterion test consisted of all 80 constructed items. Form A of the test contained one form of each item and Form B the alternate form. All

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR CRITERION TEST CORRECT RESPONSES FOR QUESTION PLACEMENT, QUESTION PACING, AND TEST TIME

Group	Experimental				Control	
	Before (n = 47)		After (n = 51)		Read only (n = 6)	Questions only (n = 5)
	Critical	Incidental	Critical	Incidental		
Immediate						
$\bar{X}$	16.28	9.62	17.80	13.09	9.36	7.91
SD	8.30	6.69	9.05	7.10	4.52	5.38
Delay						
$\bar{X}$	14.08	9.21	15.02	11.20	7.07	3.88
SD	8.09	6.22	8.61	6.89	4.96	3.11

subjects received both forms of the test, and the item order was separately randomized for both the immediate and delayed measures.

Procedures

Subjects participated in two groups (ns = 60 and 68) with subjects from all conditions present at each session. Text booklets were distributed randomly, with the restriction that about the same number of booklets from each treatment appeared at each session. Control group materials were handed out at a lower ratio to conserve n in the experimental conditions.

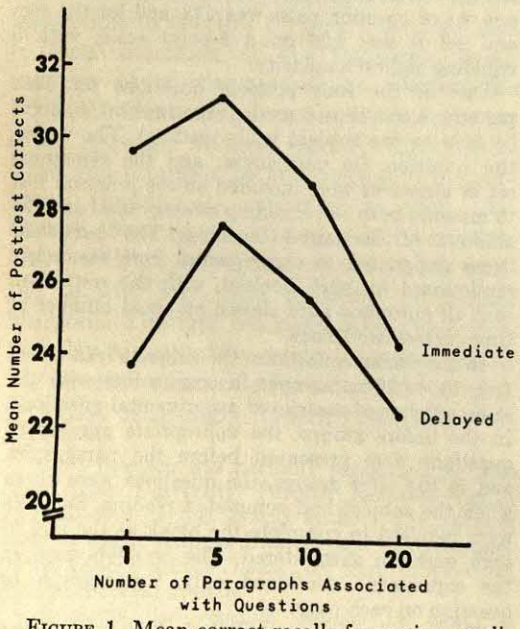


FIGURE 1. Mean correct recalls for pacing conditions on both the immediate and delayed posttests.

Each booklet contained a sheet of general instructions directing the learner to read carefully and not to refer back to material previously read. As soon as all the subjects signified that they understood the task, they were instructed to begin working through the booklet. Immediately upon completing the text, the subject raised his hand and a monitor collected his booklet, recorded his reading time, and gave him the Form A of the criterion test. When the learner had completed this form, it was collected and he received the Form B.

One week following the experimental session, the two forms of the criterion test were again administered to all participants in the same manner.

RESULTS

Table 1 contains the means and standard deviations on posttest correct responses for the experimental and control groups collapsed across the paragraph variable. A 2 (Before–After) × 4 (1, 5, 10, or 20 Paragraphs) × 2 (Critical–Incidental Item) × 2 (Immediate–Delayed Test) unweighted means analysis of variance on this data yielded significant results for the before–after ($F = 4.65$, $df = 1/90$, $p < .05$), critical–incidental item ($F = 10.69$, $df = 1/90$, $p < .01$), and immediate–delayed test ($F = 76.19$, $df = 1/90$, $p < .01$) main effects and the 1, 5, 10, or 20 Paragraphs × Immediate–Delayed Test interaction ($F = 4.66$, $df = 3/90$, $p < .01$). No other terms in this analysis reached significance. An analysis of simple effects for the 1, 5, 10, or 20 Paragraphs × Immediate–Delayed Test interaction yielded significance between immediate and delayed measures only for learners who had questions associated with individual paragraphs ($F = 4.67$, $df = 3/90$, $p < .01$). Figure 1 shows the form of the 1, 5, 10, or 20 Paragraphs × Immediate–Delayed Test interaction.

A 2 (Before–After) × 4 (1, 5, 10, or 20 Paragraphs) × 2 (Immediate–Delayed Test) unweighted means analysis of variance on the verbatim–paraphrase variable yielded no statistical significance.

A 2 (Before–After) × 4 (1, 5, 10, or 20 Paragraphs) analysis of reading times was also computed. Again, none of the terms in this analysis were significant. Clearly, posttest scores cannot be attributed to differential study time.

Discussion

The superiority of groups receiving questions after reading is consistent with available data. However, contrary to previous research, postquestions failed to differentially facilitate learning of incidental items. The fact that critical items were learned consistently better, but did not interact with other variables indicates that storage for learners of this age is best served by specific cueing devices. It may be that children disregard incidental material when cues are available to them. This finding buttresses our contention that control of inspection behaviors is a markedly different task with younger learners.

The absence of main effects for the pacing variable is consonant with earlier studies. However, the Pacing × Placement relationship found by Frase (1968b, 1968c) failed to occur. Again, this lack of agreement with our data may stem from our subjects' inability to efficiently store as much information as more experienced learners.

These data do not support our hypothesis that adjunct questions act on the type of coding in which the reader engages. With prose material, children seem able to process enough of the passage "meaning to allow recall of verbatim or paraphrased items equally well. This is an important point for further research, since there are studies which suggest that semantic development is far from complete in elementary age children (Palermo & Molfese, 1972).

The Pacing × Time of Test interaction is of interest. Over a retention interval there is a trend for forgetting to decrease as a function of the number of paragraphs seen prior to test items. However, the decrement is significant only for the subjects who had questions associated with one paragraph. Apparently, less information is lost over time when the text and associated questions are learned as larger units. Perhaps reading longer segments before testing serves to assist the subject in effectively storing more general context from the passage. Here, more intact presentation yields greater retention simply because the subject is able to answer questions with reference to the total theme of the story rather than to its specific elements. Conversely, learners receiving the fragmented presentation may learn more specific information, but this type of material has a greater chance of being forgotten. If this reasoning is sound, the most efficient use of adjunct questions would be to place them in relation to larger blocks of text.

This study suggests that adjunct questions may yield different results as a function of both the learner's age and the amount of time that passes before the posttest is given. Future research might concentrate on further assessing the degree to which learning involving adjunct devices is influenced by the age and language sophistication of the subject population.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research,* 1972, **42,** 145–170.

Frase, L. T. Learning from prose material: Length of passage, knowledge of results, and position of questions. *Journal of Educational Psychology,* 1967, **58,** 266–272.

Frase, L. T. Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Pschology,* 1968, **59,** 244–249. (a)

Frase, L. T. Some data concerning the mathemagenic hypothesis. *American Educational Research Journal,* 1968, **5,** 181–189. (b)

Frase, L. T. Some unpredicted effects of different questions upon learning from connected discourse. *Journal of Educational Psychology,* 1968, **59,** 197–201. (c)

Kulhavy, R. W., & Parsons, J. A. Learning-criterion error perseveration in text materials. *Journal of Educational Psychology,* 1972, **63,** 81–86.

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review,* 1956, **63,** 81–97.

Palermo, D. S., & Molfese, D. L. Langage acquisition from age five onward. *Psychological Bulletin,* 1972, **78,** 409–428.

Rothkopf, E. Z., & Bisbicos, E. E. Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology,* 1967, **58,** 56–61.

Sachs, J. S. Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics,* 1967, **2,** 437–442.

(Received February 22, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 216-224

LEARNING BY OBSERVING VERSUS LEARNING BY DOING[1]

DOUGLAS K. CHALMERS[2]
University of California, Irvine

MILTON E. ROSENBAUM
University of Iowa

Differences between performers and observers were investigated in a concept-transfer task. The method involved assessing the transfer due to reversal, nonreversal, and irrelevant or control shifts as a function of either performing or observing during an initial training task. The hypothesis that conceptual responses are readily learned by observation was confirmed. When error rates of performers and observers were compared on the shift tasks, it was found that (*a*) performers, but not observers, showed negative transfer on the reversal-shift task and (*b*) observers showed a smaller overall error rate than performers. These effects were consistent with the second hypothesis, which posited a relatively reduced degree of associative interference as a function of observational training experience.

Experimental research on learning as a function of the observation of the behavior of other persons has emphasized the replication of the specific responses the model has been observed to perform. Little attention has been given to the transfer or generalization of the observed experience to new situations, despite the acknowledgment in discussions of observational learning of the importance of this process in social development (cf. Aronfreed, 1969; Bandura, 1969; Rosenbaum & Arenson, 1968).

The present study purports to examine some aspects of the differences between direct participation and observation in concept learning and transfer. For this kind of comparison to be made adequately, it is essential that the experimental design include a Task 2 measure of the behavior of the Task 1 *performer* as well as that of the observer. Only a few studies (but none including concept-learning tasks) have included this type of comparison (cf. Bruning, 1965; Riopelle, 1960; Rosenbaum, 1967).

The technique employed in the present study involves a comparison of the transfer effects of reversal and nonreversal shifts as a function of performing or observing the initial task. Unlike studies of incidental learning by observation (Bandura & Huston, 1961; Berger, 1961), the observer was initially instructed that he would later perform and that the observed task would be relevant to his later performance. A reversal shift typically involves learning two successive sorting tasks with correct responses for Task 2 being based on the same dimension of the stimuli as on Task 1 but with the subject being required to re-pair the previous sorting responses with the dimension values. Nonreversal shifts require learning a second task based on some dimension of the stimuli which was present but irrelevant on Task 1. For example, if Task 1 required sorting cards into certain categories on the basis of shape, a reversal shift would require a new sorting scheme also based on shape, whereas a nonreversal shift might involve one based on the color of the cards.

Previous investigations with the concept-shift task have employed performers only. Experiments with college students as performers have consistently found that reversal shifts are learned more quickly than nonreversal shifts (cf. Wolff, 1967). Kendler and Kendler (1968) interpret this effect by employing mediational constructs in a stimulus–response framework. By assuming that mediating conceptual responses have

[1] Partial support for this study was provided by a grant from the National Science Foundation to the second author.

[2] Requests for reprints should be sent to Douglas K. Chalmers, Program in Language and Behavior, School of Social Sciences, University of California, Irvine, California 92664.

been acquired on Task 1, a reversal shift requires fewer new associations than does a nonreversal shift. In a study by Isaacs and Duncan (1962), it was found that a reversal group showed Task 2 learning inferior to that of a control group in which Task 1 consisted of an entirely different set of stimulus dimensions from those relevant on Task 2. They reasoned that the subjects performing a reversal shift must extinguish Task 1 *specific* associations between stimuli and task responses before Task 2 learning can occur, whereas this was not the case for subjects in the control group. The present study also made use of this type of control condition, in which the specific effects of Task 1 were irrelevant to Task 2.

In extending Miller and Dollard's (1941) analysis of imitation to observational-learning situations, some writers have emphasized the role of covert imitative responses occurring during observation (Aronfreed, 1969; May, 1946; Rosenbaum & Arenson, 1968). Having learned in the past that matching the responses of others frequently leads to rewards, children acquire a generalized tendency to imitate, which may be expressed without the necessity of overt matching behavior. Learning will occur as a result of observation when implicit or covert matching responses come to be elicited by the appropriate situational stimuli alone. The covertly learned responses can then be utilized in later contexts.

It is assumed that both performers and observers covertly respond to the relevant stimuli during learning. Furthermore, the effect of the performer's overt response is seen simply as adding another element to his total response. It is held, therefore, that while the total response of the performer consists of both covert and overt components, the response of the observer consists of the covert component alone.

In providing a test of transfer of both *specific* stimulus–response associations and implicit *conceptual* responses, the reversal–nonreversal technique affords distinct advantages for studying observational learning. Conceptual responses are, of course, covert for both performers and observers. On the assumption that covert responses are learned as readily by observers as by performers, it is predicted that the usual finding of superiority of a reversal over a nonreversal shift will hold equally well for observers.

In regard to specific associations, on the other hand, whereas observers are assumed to have acquired covert associations alone, performers have acquired both overt and covert associative components. It is further predicted, therefore, that observers will undergo less specific interference from their initial (Task 1) associations than will performers on the shift tasks.

Method

Materials

Materials were identical to those employed by Walther (1962). The stimuli were slides showing figures varying in three dimensions. For reversal and nonreversal groups on Task 1 and for all subjects on Task 2, the dimensions were shape (square, circle, diamond, and cross), number of figures (1, 2, 3, and 4), and location of figures (top, bottom, right, and left). All figures were black on a white background. For the control groups on Task 1, the figures, all of which were square, varied in color (red, green, blue, and yellow), size (small, medium-small, medium-large, and large), and location (top, bottom, right, and left).

In order to eliminate continued reinforcement of Task 1 responses on Task 2 for the nonreversal groups, certain combinations of shape and number never appeared together. These were square and 1, circle and 2, diamond and 3, and cross and 4. Due to the exclusion of these combinations, at the beginning of Task 2 there was no associative strength between the values of the relevant stimulus dimension and the correct responses for either the reversal or the nonreversal groups. After eliminating these combinations, a total of 48 stimuli remained to be presented to these groups in Task 1.

To equate the number of stimuli on Task 1 for reversal, nonreversal, and control groups, certain combinations of color and size were eliminated from the Task 1 stimuli presented to the control groups. These were red and small, green and medium-small, yellow and medium-large, and blue and large. In all conditions, the four locations occurred equally often with each value of shape, number, and color.

The stimuli, having been photographed on 16-millimeter film, were mounted on 2 × 2 inch, 16-millimeter slide mounts for projection by a Kodak Carousel slide projector. For all sets of Task 1 stimuli, two sequences of the 48 slides were made, such that each slide appeared once in each sequence and assignment to position in the

sequence was based on a table of random numbers. The only restriction on the random order was that no value of any dimension should immediately follow itself. Since the slide drum for the Carousel projector holds a maximum of 80 slides, the last 16 slides in the series were eliminated. This left a continuous series of 80 slides for Task 1.

Apparatus

Totally separated by a partition of black plywood, the two subjects (in the observe condition) were seated in such a way that the confederate was invariably to the left side of the naive subject. The subjects faced a milk-glass screen (36.5 centimeters high and 42 centimeters wide), mounted on a fiberboard partition which was painted flat black and placed on a table. This partition served to shield the experimenter from the view of the subjects. The screen was at eye level and equidistant (1.1 meter) from each subject. In the perform condition, only the naive subject was present, and he was always seated on the same side of the partition as the naive subjects in the observe condition.

The slides, projected on the back of the milk-glass screen, were approximately 15 centimeters square. A green and a red jewel light were centrally located 7.5 centimeters and 15 centimeters, respectively, above the screen.

Responses were four syllables selected from the Glaze list of 40% association value syllables: PEX, KUG, MOF, and WIJ (Hilgard, 1951). Each of the four nonsense syllables, printed in large black letters, was located at a different corner of a white 8 × 11 inch pasteboard card. The card was taped to the fiberboard 1.25 centimeters directly below the milk-glass screen. This card was available throughout the session for reference. Response familiarization was given as part of the instructions.

On the appropriate side of the plywood partition separating the subjects, there was a printed letter, "A" for the confederate and "B" for the naive subject. The experimenter used these letters in referring to the subjects in the observe condition. Attached to the underside of the confederate's chair, hidden from view, was a metal container which made available to the confederate his response card (a 5 × 8 inch note card) at the beginning of each observe session. The particular response sequences on the cards were prearranged by the experimenter, as described below.

The subject responded to the stimuli by saying aloud one of the four syllables. If the response was correct, the green light came on. If the response was incorrect, the red light came on. In either case, the light appeared as soon after the subject's response as the experimenter could press his signal switch and remained on for approximately three seconds. Immediately after the signal light went off, the experimenter presented the next slide.

Design

To increase the precision and efficiency of the experiment, a confederate (an advanced undergraduate psychology major) was employed throughout the observe condition of the experiment. This procedure served a threefold purpose. Most importantly, by employing a confederate as the Task 1 performer in the observe condition, it was possible to exactly match each Task 1 performance in the observe group with a Task 1 performance in the perform group. Thus it was insured that a comparison of Task 2 performance between perform and observe groups would be independent of both the number of trials to Task 1 criterion and the particular learning sequences involved in the attainment of the Task 1 concept. Second, the addition of a confederate in the observe condition made certain that each of the Task 1 responses was quite clear and audible for the observer's reception. Third, the fact that the same person (the confederate) always served as Task 1 performer in the observe group insured a constant stimulus situation for this group.

In the perform group on Task 1, an equal number of subjects learned a shape, number, and color concept to a criterion of 10 consecutive correct responses. On Task 2, half of the subjects in each group learned a shape concept and half a number concept. Those who learned a shape concept on Task 1 and a shape concept on Task 2, with re-pairing of concept instances and responses on the second task, constituted a reversal-shift group, together with those who received number–number training. Shape–number and number–shape subjects constituted a nonreversal-shift group and color–shape and color–number subjects constituted a control-shift group.

Exactly the same design applies to the observe group, except that the confederate performed Task 1, while the naive subject simply observed his partner's task performance. At the completion of Task 1, the Task 1 observer then performed Task 2, while the confederate remained silent. Naive subjects were led to believe that the confederate was also naive. Each Task 1 performance of the confederate was a reproduction of the Task 1 performance of a subject in the perform group, thus providing past experience for each observer equivalent to that of a performer in the perform group.

On Task 2, half of the subjects in each of the six conditions learned a number concept and half learned a shape concept. When the number concept was relevant, the subject was to learn to say PEX to the presentation of any three figures, regardless of their shape or location, WIJ to four figures, KUG to one figure, and MOF to two figures. The correct pairings of the shape–stimulus values and nonsense syllables on Task 2 were diamond–PEX, cross–WIJ, square–KUG, and circle–MOF.

The nonreversal group consisted of two subgroups determined by which concept (number or shape) was learned first. One subgroup (number-

shape) learned the number concept first, then the shape concept in Task 2. The other nonreversal subgroup received the shape–number order. The control group was presented one color concept as Task 1 and received either number or shape as Task 2.

In order to increase the generality of the findings, three different combinations of values and responses were used in the reversal conditions on Task 1 for both shape (S_1, S_2, S_3) and number (N_1, N_2, N_3). This was done to anticipate the possibility that some within-dimension shifts might be easier than others. Thus the reversal condition consisted of six equally allocated Task 1 to Task 2 sequences: S_1–S, S_2–S, S_3–S and N_1–N, N_2–N, N_3–N. Although not all of the reversal shifts employed involved "literal" reversals (shifts that are made within pairs of Task 1 associations), the authors follow Harrow and Buchwald (1962) in maintaining the "reversal" terminology for any shift that is made on the basis of the same dimension as was relevant on Task 1. Harrow and Buchwald, finding no differences between literal and nonliteral reversal shifts, argue that a sufficient condition for a reversal shift is the common relevance of the same dimensional values on both tasks.

Procedure

During the three weeks of testing, perform sessions were run on the first half and observe sessions on the second half of each week in roughly equal numbers over shift conditions. This alternation procedure was necessary in order to obtain groups matched on Task 1 performance. Aside from this consideration, the subjects were assigned at random to the reversal, nonreversal, or control conditions. Upon arrival for testing, the subjects were seated and read the instructions. The experimenter answered any questions that had to do with procedure but gave no information concerning the nature of the figures or the classes into which they were to be divided.

In the perform condition, if the subject reached the Task 1 criterion of 10 consecutive correct responses within 100 trials, exclusive of criterion trials, he was shifted to Task 2. For the subjects in the observe condition, it was necessary to pause long enough to notify the observer that he was now to perform. Therefore, a 30-second interval was introduced between Tasks 1 and 2 in *both* perform and observe groups. In an attempt to reduce the possible confounding effect of differential sets, the subjects in both groups were informed during the initial instructions and at the completion of Task 1 that the purpose of the pause was to "change the slide tray." The subjects continued on Task 2 until they made 10 consecutive correct responses. Any subject who did not reach this criterion within 100 trials, exclusive of criterion trials, was given a score of 100 and the trials were terminated.

Each Task 1 performer started at the beginning of the sequence of 80 slides. Task 2 was initiated just after that point in a sequence corresponding to the number of slides used in Task 1. The subjects matched in the perform and observe conditions were given identical slide sequences.

In an attempt to discover those subjects in the observe condition who did not learn Task 1 and to compare performer–observer retention, a brief written questionnaire consisting of two questions was administered to all subjects immediately following completion of Task 2. One question was phrased to elicit recall of the Task 1 concept and the second to test recall of the correct Task 1 pairings among the values of the relevant dimension and the nonsense syllables.

Subjects

Subjects were 100 male students from the undergraduate introductory psychology course at the University of Iowa who were given course credit for participation. Any subject who did not learn Task 1 to the required criterion within 100 trials, exclusive of criterion trials, was dropped from the experiment. Four subjects were eliminated for this reason. Two of these subjects were to have learned a color concept, one a number concept, and one a shape concept.

All subjects in the control group were administered the Dvorine pseudo-isochromatic plates (color vision test) at the end of the experiment. Any subject who failed to respond correctly on more than three plates was eliminated from the study. Four subjects were dropped for this reason.

Results

Task 1

The mean number of trials required to learn Task 1 for subjects who were to be exposed to the various shift conditions, exclusive of criterion trials, was 40.0, 32.4, and 38.2 for the reversal, nonreversal, and control groups, respectively. The mean number of trials for training in Task 1 on the various concept dimensions was 32.5, 39.9, and 38.2 for the shape, number, and color concepts, respectively. Separate analyses of variance revealed no differences among the means of either analysis. Both F ratios were of a magnitude smaller than 1.

Task 2: Trials to Criterion

Seven subjects failed to learn Task 2 to the required criterion within 100 trials. These subjects were given a score of 100, and Task 2 was terminated. Four of these subjects served in the observe condition and three in the perform condition. Of the observers, three were in the nonreversal condition and one in the reversal condition. Of the

TABLE 1
MEAN NUMBER OF TRIALS TO LEARN TASK 2

Type of shift	Task 1 experience				
	Perform		Observe		Combined
	M	*SD*	*M*	*SD*	*M*
Reversal	21.1	13.7	29.8	23.6	25.4
Nonreversal	48.3	26.9	53.3	29.1	50.8
Irrelevant	19.3	24.5	24.8	18.2	22.1
Combined	29.6		35.9		

performers, two were in the nonreversal condition and one in the control condition. Table 1 presents the mean number of trials required to learn Task 2, exclusive of criterion trials, and the standard deviations for each subgroup. Because inspection of these data revealed heterogeneity of variance and some departure from normality, a more conservative level of significance was adopted ($p < .025$).

Although observers required a slightly greater number of trials to criterion than performers in all shift conditions, there was no reliable difference between these two conditions ($F = 2.0, df = 1/102$). Type of shift was significant ($F = 17.8, df = 2/102, p < .01$). The critical difference (i.e., the minimum difference required for significance; Lindquist, 1953), computed for the row means with a value of t significant at the .025 level, was 11.8. A difference of this size occurred in the direction of a greater number of trials to criterion for the nonreversal group than for the reversal and control groups, which did not differ from each other. The finding of no interaction between Task 1 experience and type of shift ($F < 1.0, df = 2/102$) indicated that the differences due to type of shift were of the same magnitude for both performers and observers.

Errors

To reveal some of the details of Task 2 performance, it is of value to employ a measure that assesses the relative occurrence of errors on the shift tasks independently of the total number of trials to criterion. A measure suitable for this purpose was the ratio of errors to the total number of responses for each subject.

The mean ratio of errors to total responses (excluding criterion trials) for all conditions is presented in Table 2. An analysis of variance of these data revealed main effects due to type of shift and Task 1 experience ($F = 17.7, df = 2/102, p < .01$ and $F = 9.1, df = 1/102, p < .01$, respectively). Column means showed a lower mean error ratio for observers; row means showed the lowest mean error ratio for subjects in the control condition. However, the additional finding of significant interaction effects ($F = 7.5, df = 2/102, p < .01$) qualifies a direct interpretation of the main effects. The critical difference of .079 (computed for $p = .025$), applicable to the six cells of the table, warrants the following statements. Performers obtained a significantly smaller error ratio in the control condition than in the reversal and nonreversal conditions. Observers, on the other hand, produced a smaller error ratio in both the reversal *and* control conditions than in the nonreversal condition. Perhaps of greatest interest is the finding that observers produced a significantly smaller error ratio than performers in the reversal condition but did not differ reliably from performers in the nonreversal or control conditions. Each mean of Table 2 could, of course, be converted into a mean proportion of correct responses by subtracting the error ratio from 1.

Trials to successive criteria

On the basis of the error ratio measure used above, two events occurred of special interest that were not accounted for in terms of the trials-to-criterion measure. A reversal shift brought about a greater pro-

TABLE 2
MEAN RATIO OF ERRORS TO TOTAL RESPONSES

Type of shift	Task 1 experience		
	Perform	Observe	Combined
Reversal	.693	.530	.612
Nonreversal	.698	.654	.676
Irrelevant	.517	.542	.529
Combined	.636	.575	

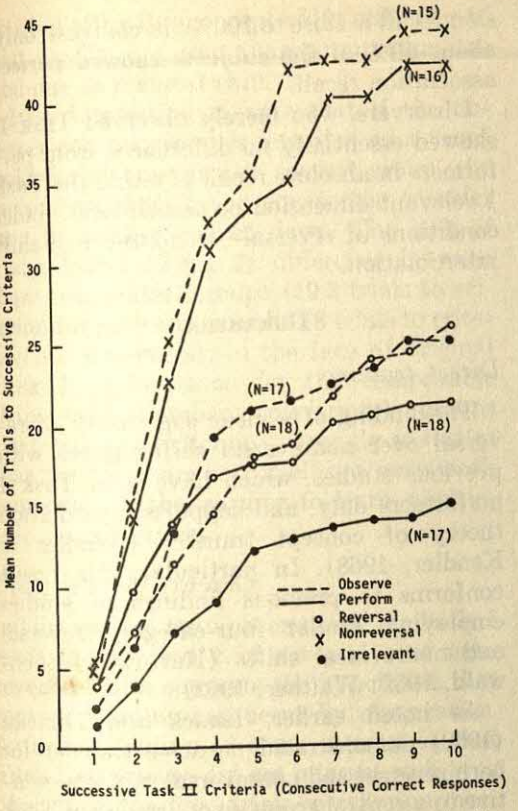


FIGURE 1. Mean number of trials to successive criteria during Task 2.

portion of errors for Task 1 performers than for Task 1 observers. A reversal shift in the perform condition also brought about a greater proportion of errors than did the control shift. These findings suggest that although these groups did not differ with respect to the attainment of a particular criterion, differences did indeed occur during the process of reaching that criterion.

Figure 1 shows the number of trials to criterion (exclusive of criterion trials) as a function of criterion difficulty. The criteria employed along the abscissa are each in terms of number of consecutive correct responses. To plot these curves it was necessary to eliminate the seven subjects who failed to reach the Task 2 criterion of 10 consecutive correct responses. Since the purpose of the curves in Figure 1 is to describe concept acquisition, the inclusion of these subjects would be inappropriate. In addition, since the sample of subjects in the perform condition was restricted to only those subjects who were able to learn Task 1, elimination of all nonlearners from Task 2 provided a comparable screening device for observers.

Due to the nature of the trials-to-successive-criteria measure and to the lack of cell proportionality, no statistical analyses were performed on the data of Figure 1. Furthermore, these data are intended merely as a descriptive commentary on the error rate data discussed above. The curves related to error rate as follows: (*a*) a higher average slope reflects a greater number of errors between the attainment of successive criteria; (*b*) positive acceleration is produced either by an increase or by no change in the number of errors between criteria; (*c*) both linearity and negative acceleration are brought about by a reduction in the number of errors between criteria; and (*d*) leveling of the curves occurs where the error rate is reduced to zero (e.g., a jump from 6 consecutive correct responses to 10 consecutive correct).

The most apparent observation from Figure 1 is the relatively high acquisition slopes for the nonreversal groups, indicating much slower progress toward successive criteria throughout. In addition, a higher criterion is required for nonreversal subjects before any appreciable error reduction (negative acceleration) takes place.

The finding of negative transfer for the error rate measure for reversal performers with respect to the control performers appears to be consistently the case at all criterion levels of Figure 1 (solid lines in lower portion of figure). The associated finding of no error rate difference for reversal and control observers is shown by the broken lines of Figure 1, where observers exhibit either positive or essentially zero transfer throughout.

Finally, it is of interest to compare performer–observer differences in reversal-shift performance. The reliable finding of a greater error rate for performers than observers is reflected in the open-dot curves of Figure 1 by two observations: (*a*) a greater absolute number of errors for performers than observers at the lower criteria and (*b*) a faster rate of error reduction be-

TABLE 3

NUMBER OF SUBJECTS SHOWING RECALL OF THE RELEVANT TASK 1 DIMENSION AND OF ALL THE TASK 1 STIMULUS–RESPONSE ASSOCIATIONS

Type of shift	Task 1 experience			
	Perform		Observe	
	Dimension	Association	Dimension	Association
Reversal	16	10	16	10
Nonreversal	18	8	17	11
Irrelevant	18	12	13	8

Note. n = 18 per group.

tween criteria for performers than for observers. Observers appear then to have the edge on performers, in terms of fewer errors, only in the beginning phases of reversal transfer.

In comparing performers with observers, the curves strongly suggest divergence in the control condition (solid dots of Figure 1). As the control shift was irrelevant to Task 2 requirements, this divergence is likely due to performer–observer differences in the transfer of nonspecific factors. That is, the positive advantage for performers over observers of such factors as visual–vocal coordination and temporal adjustment appears to increase as the difficulty of the transfer criteria increases.

Recall

Following the shift task, the subjects were administered a questionnaire in order to test recall of Task 1 training. The first item, which stipulated that the subject was to "explain in general terms how the slides were classified . . . before the pause," tested recall of the Task 1 relevant dimension. Recall of the specific Task 1 stimulus–response pairings was tested by asking the subject to "try to recall which nonsense words were associated with which slide categories before the pause."

Table 3 shows the number of subjects (n = 18 per group) in all 12 conditions who showed correct recall of the Task 1 relevant dimension and of the Task 1 associations. For performers, recall is uniform across type of interpolated shift. Dimension recall is close to 100%; in contrast, only about 50% of the subjects showed perfect association recall.

Observers, who merely observed Task 1, showed essentially no differences from performers in absolute recall of either the Task 1 relevant dimension or associations, under conditions of reversal- or nonreversal-shift interpolation.

DISCUSSION

Direct training

The finding of a clear superiority of reversal over nonreversal shifts agrees with previous studies, which have used Task 1 performers only, and supports a mediation theory of concept transfer (Kendler & Kendler, 1968). In particular, this result conforms to previous findings of studies employing similar four-category reversal and nonreversal shifts (Harrow & Buchwald, 1962; Walther, 1962).

As noted earlier, Isaacs and Duncan (1962) demonstrated negative transfer for both reversal and nonreversal subjects, relative to a control group given irrelevant Task 1 training. In terms of the trials-to-criterion measure, the present study revealed negative transfer for the nonreversal group only, as did the Walther (1962) study. However, when a more discriminating error measure was employed for performers' data, negative transfer was indicated for reversal subjects, as well as for nonreversal subjects. On a specific stimulus-overt response association level, then, reversal and nonreversal performers show negative transfer of Task 1 associations, as expected from interference theory in verbal learning (Postman, 1961). Such low-level interference occurs despite the large concept-transfer effect, as indicated by the reversal–nonreversal differences.

Finally, the reversal–nonreversal concept-transfer effect must be attributed to negative transfer of the nonreversal group, as opposed to positive transfer of the reversal group. That is, since transfer for reversal subjects was either zero (for trials to criterion) or negative (for error rate) with respect to the control group, the inferior performance of nonreversal subjects must be

attributed to the concept-shift difficulty of that condition rather than to any facilitation due to reversal shift.

It is interesting to note that if the observer control group is regarded as a "control" for the performer control group, the factors responsible for nonspecific transfer may be further broken down. The lack of a significant (Task 2) difference between these two control groups (19.3 trials to criterion for performers and 24.8 trials to criterion for observers), in the face of original Task 1 performance for the comparable shape and number (performer) groups (39.6 trials to criterion), strongly indicates that "performance set" itself is a very weak component of the learning-to-learn, nonspecific transfer effect.

Observational training

The finding that observation of Task 1 leads to superiority of reversal over nonreversal shifts supports the hypothesis that concept learning can occur by observation. A further finding of some interest is that Task 1 experience did not interact with type of shift on the final trials-to-criterion measure. That is, the difference between reversal and nonreversal groups for observers was of the same magnitude as that for performers. Conceptual responses acquired by observation, then, appear to be no less potent than those acquired by direct performance.

Although observers and performers did not differ on the final trials-to-criterion measure, observers showed a reliably smaller overall error ratio than performers on the transfer task. In addition, their patterns of error rates differed. While performers, as noted above, showed negative transfer for both reversal and nonreversal shifts relative to their control-shift error rate, observers did not show this negative transfer in the reversal condition. Moreover, the smaller error rate for observers was primarily attributed to the reversal-shift condition. The second prediction, then, that observers would undergo less associative interference, was confirmed.

To summarize, the error analysis revealed that relative to performers, observers demonstrated an immunity to perseveration of Task 1 associations, especially at the beginning of transfer, where specific interference effects should be greatest. This evidence is viewed as support for the thesis that responses made by performers are more resistant to extinction than are the covert counterparts of these responses made by observers. More precisely, whereas a performer must extinguish both the covert and the overt components of his response, an observer need only extinguish a covert response. An alternate interpretation is that observers did not learn the Task 1 associations as well as did the performers and, therefore, experienced less interference as a result. The finding that observers did not differ from performers in overall recall of these associations would not favor the latter interpretation.

In conclusion, although concept attainment occurs quite as readily under observational as under direct training conditions, observational training appears to provide a certain "flexibility" in the attained concept. This "flexibility" is reflected in a detachment of the concept from its specific learning conditions.

REFERENCES

Aronfreed, J. The problem of imitation. In L. P. Lipsitt and H. W. Reese (Eds.), *Advances in child development and behavior*. Vol. 4. New York: Academic Press, 1969.

Bandura, A. Social-learning theory of identificatory processes. In D. A. Goslin (Ed.), *Handbook of socialization theory and research*. Chicago: Rand McNally, 1969.

Bandura, A., & Huston, A. C. Identification as a process of incidental learning. *Journal of Abnormal and Social Psychology*, 1961, **63**, 311–318.

Berger, S. M. Incidental learning through vicarious reinforcement. *Psychological Reports*, 1961, **9**, 477–491.

Bruning, J. L. Direct and vicarious effects of a shift in magnitude of reward on performance. *Journal of Personality and Social Psychology*, 1965, **2**, 278–282.

Harrow, M., & Buchwald, A. M. Reversal and nonreversal shifts in concept formation using consistent and inconsistent responses. *Journal of Experimental Psychology*, 1962, **64**, 476–481.

Hilgard, E. R. Methods and procedures in the study of learning. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951.

Isaacs, I. D., & Duncan, C. P. Reversal and nonreversal shifts within and between dimensions in

concept formation. *Journal of Experimental Psychology,* 1962, **64,** 580–585.

KENDLER, H. H., & KENDLER, T. S. Mediation and conceptual behavior. In K. W. Spence and J. T. Spence (Eds.), *Psychology of learning and motivation.* Vol. 2. New York: Academic Press, 1968.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

MAY, M. A. The psychology of learning from demonstration films. *Journal of Educational Psychology,* 1946, **37,** 1–12.

MILLER, N. E., & DOLLARD, J. *Social learning and imitation.* New Haven, Conn.: Yale University Press, 1941.

POSTMAN, L. J. The present status of interference theory. In C. N. Cofer (Ed.), *Verbal learning and verbal behavior.* New York: McGraw-Hill, 1961.

RIOPELLE, A. J. Observational learning of a position habit by monkeys. *Journal of Comparative and Physiological Psychology,* 1960, **53,** 426–428.

ROSENBAUM, M. E. The effect of verbalization of correct responses by performers and observers on retention. *Child Development,* 1967, **38,** 615–622.

ROSENBAUM, M. E., & ARENSON, S. J. Observational learning: Some theory, some experiments, some data. In E. C. Simmel, R. A. Hoppe, & G. A. Milton (Eds.), *Social facilitation and imitative behavior.* Boston: Allyn-Bacon, 1968.

WALTHER, W. E. Degree of Task I learning and reversal and nonreversal shifts in concept formation. (Doctoral dissertation, University of Iowa) Ann Arbor, Mich.: University Microfilms, 1962, No. 63–985.

WOLFF, J. L. Concept-shift and discrimination-reversal learning in humans. *Psychological Bulletin,* 1967, **68,** 369–408.

(Received March 12, 1973)

ERRATUM

In the article "Effect of Anxiety, Response Mode, Subject Matter Familiarity, and Program Length on Achievement in Computer-Assisted Learning" by Barbara L. Leherissey, Harold F. O'Neil, Jr., Darlene L. Heinrich, and Duncan N. Hansen which appeared in the June 1973, Volume 64 issue of the *Journal of Educational Psychology,* two of the four groups in Figure 6 were incorrectly labeled: The Constructed Response-Long group should have been labeled Constructed Response-Short and the Constructed Response-Short group should have been labeled Constructed Response-Long. In addition, the text that refers to Figure 6 should be changed to reflect this correction. These changes do not affect the remaining results and discussion sections.

Journal of Educational Psychology
1974, Vol. 66, No. 2, 225–232

EFFECT OF TYPE OF OBJECTIVE, LEVEL OF TEST QUESTIONS, AND THE JUDGED IMPORTANCE OF TESTED MATERIALS UPON POSTTEST PERFORMANCE

ORPHA K. DUELL[1]
Wichita State University

The purpose was to discover under what conditions providing behavioral objectives during study improves the amount learned. College seniors (n = 80 and n = 87) participated in two separate experiments. The findings support the hypothesis that providing students with behavioral objectives during study produces greater learning only if the objectives direct the student to learn information which he would not classify as important or likely to be tested.

Some investigations which have studied the effects on learning of providing students with behavioral objectives as they study text materials have shown learning advantages for students provided behavioral objectives (Morse & Tillman, 1972; Rothkopf, 1972; Rothkopf & Kaplan, 1972; Webb & Cormier, 1972), while others do not (Jenkins & Deno, 1971; Merrill & Towle, 1972; Oswald & Fletcher, 1970; Yelon & Schmidt, 1971). Because of the mixture of results, it would seem that providing behavioral objectives is advantageous under some circumstances but not all. A variable that might be a controlling factor is the test expectations which subjects bring to the learning situation. A student's test expectations would influence what is learned by determining what parts of the text the student chooses to actively process and what he ignores. This assumes that as long as the teacher is not providing assignments, tests, and/or questions which force the active processing of materials, the student either consciously or unconsciously chooses what he will actively process.

Student expectations would logically be influenced by factors such as past experience with this teacher, this type of course, and this type of text, as well as what type of test the teacher says he will give and the type of test previous students say this teacher gives.

One characteristic of the textual material which might affect test expectations and, therefore, the effectiveness of behavioral objectives is the agreement between the emphasis placed upon information within the text and what is included in the behavioral objectives. For example, consider a passage that introduces four concepts, one is expanded upon through examples and descriptions, while the others are barely mentioned in passing. If the behavioral objective concerns the concept with which the majority of the passage deals, the behavioral objective tells the student nothing more than is already indicated by the reading passage; however, if the behavioral objective includes the major concept plus one of the other three concepts, the student may benefit from having the objective, since it provides guides that differ from the ones provided by the text.

Yet another factor that may affect the learner's test expectations is his past experience with tests. Reviews of teacher-made tests indicate that most of the items (90%–95%) require rote recall of specific information (Thorndike & Hagen, 1969), despite the fact that most teachers would agree that "higher level" behaviors, as defined by Bloom (1956), are more important goals. This would lead to the prediction that students not provided with orienting directions such as behavioral objectives would probably expect to be given recall questions and would not expect higher level questions over text material.

[1] Requests for reprints should be sent to Orpha K. Duell, College of Education, Wichita State University, Wichita, Kansas 67208.

Experiment 1

The first study was designed to investigate the joint effects of the level of the test questions and the availability of behavioral objectives on learning from written prose. The prediction was that the difference between the behavioral objective and nonbehavioral objective groups would be nonsignificant for recall questions but significant for application questions. Each reading passage was designed to introduce three related concepts. All concepts were defined and expanded by at least one example.

Method

Subjects

Fifty-six college seniors enrolled in two sections of an educational psychology course taken just before student teaching were randomly assigned to the two experimental conditions so that approximately half of each section was in each of the experimental conditions. Each experimental condition contained 28 subjects. An additional group of 24 students in another section of the same course acted as a control group.

Materials

Passages. Five passages were written describing the processes of prompting, shaping, using negative reinforcement, using advanced organizers, and measuring test-retest reliability. These topics were chosen because they were relevant to the course in which the students were enrolled.

Each passage contained approximately 410 words and all passages were similar in organization. The first paragraph contained a descriptive example of the process. The second paragraph introduced the term, its definition, and the names of two psychologists associated with the process. The names were chosen so that each name was used only once. The third paragraph consisted of a second descriptive example of the process. The fourth and final paragraph contained two related technical terms, their definitions, a descriptive example of each of them, and a summary sentence. Each time a technical term was introduced for the first time it was underlined.

Posttest. Multiple-choice items with four alternatives were prepared from the content of each of the five reading passages. The test items and the letter of the correct choice were randomly ordered.

Two levels of test questions were used. Recognition questions were questions which required the student to recognize something that he had read in the passage. Three recognition questions were constructed for each reading passage. Two required the student to choose definitions. The terms to be defined were the major term introduced by the passage and one of the two related technical terms from the fourth paragraph. The related term was randomly chosen so that for two of the passages the first term introduced was tested, while the second term was tested for the other three. The third recognition question named the process described in the passage and asked the student to choose the psychologist's name associated with it. Since two psychologists were named in each passage, the name to be tested was randomly chosen so that the first name was tested for three of the passages and the second name for the other two.

Three application questions were prepared for each reading passage. In each, the student chose the situation which was an example of the process named in the question. All alternatives described situations different from those in the reading passages. All three questions dealt with the major process introduced in the reading passage.

Objectives. The behavioral objectives were written in the style described by Mager (1962), except they contained no minimal standard. The following is an example objective and a portion of the preceding directions:

> When you have finished studying the passages you will be given a 30-item multiple-choice test to determine if you can do the following activities:
>
> 1. Choose from among four example situations (that are different from those described in the reading passages) the one that is an example of a specified process. In each question the specified process will be one of the following:
>
> advanced organizers
> negative reinforcement
> prompting
> shaping
> test-retest reliability.

Two additional objectives were included, one for the items on psychologist's names and one for the items requiring definitions. Their form was similar to that used in the example.

Students given nonbehavioral objectives were told, "you will be given a 30-item multiple-choice test to determine if you know and understand the material in the reading passages." This nonbehavioral objective was thought to be similar to what students are frequently told or is inferred by them when given reading assignments to be covered by an exam.

Objective use questionnaire. A questionnaire was prepared to find out whether the students given the behavioral objectives thought the test would be based on the objectives and whether they actually used the objectives to aid them in their study. If the student did not use the objectives during study, there would be no reason to expect the predicted benefit on application questions.

Procedure

Students came to class at the regular hour with no prior knowledge of the experiment. They were

given manila envelopes bearing their names. Each packet contained the reading passages and one of the two sets of directions. Both sets of directions told the student that he would be held responsible for the materials for the class in which he was enrolled, that he could use as much of the two-hour class period as he needed, that he could make marks on the reading passages, and that he could refer back to the directions during study. These instructions were followed by either the behavioral or nonbehavioral objectives, depending on the condition to which the student was randomly assigned. Within each condition, the five reading passages appeared in three different random orders. The students were asked to record the time on their green time sheets before they began the directions, after they finished the directions, and after they finished studying the reading passages. The time was written on the blackboard and it was changed every minute.

When a student indicated that he was finished studying, the study materials were picked up and the student was given a manila envelope containing a copy of the test. If the student was in the behavioral objective group, the objective use questionnaire was attached to the back of the test.

The control group was given the test and told to do their best even though they had not read the passages the test questions covered.

Results

Test Data

Table 1 contains the means and standard deviations for the experimental groups and the control group subdivided by level of question.

The experimental groups—behavioral and nonbehavioral objective groups combined ($M = 23.55$, $SD = 4.36$)—did perform significantly better ($t = 10.17$, $df = 78$, $p < .01$) than the control group ($M = 13.58$, $SD = 3.86$) on the posttest, indicating that learning had occurred in the experimental setting.

Three orthogonal planned comparisons were completed according to the procedure outlined in Hays (1963) but with one exception. Since the data involved repeated measures, the mean-square interaction term from a two-way analysis of variance was used in calculating the *t*s instead of the mean square within.[2] The first factor in the 4 × 28 analysis of variance consisted of the four different experimental conditions, while the second factor consisted of levels, where each level was a pair of subjects, one from the nonbehavioral objective group randomly paired with one from the behavioral objective group.

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR THE POSTTEST SCORES IN EXPERIMENT 1

Group/question	*M*	*SD*
Behavioral		
Recognition	13.07	2.11
Application	11.07	3.02
Nonbehavioral		
Recognition	11.07	2.62
Application	11.89	2.17
Control		
Recognition	6.58	2.10
Application	7.00	2.30

Contrary to the predictions, two-tailed tests revealed there was a nonsignificant difference between the behavioral and nonbehavioral objective group on the application questions ($t = 1.42$, $df = 81$, $p > .05$) and a significant difference on the recall questions ($t = 3.45$, $df = 81$, $p < .01$), with the behavioral objective group performing significantly better than the nonbehavioral objective group as can be seen in Table 1. The difference between everyone on the recognition questions and everyone on the application questions was nonsignificant ($t = 1.44$, $df = 81$, $p > .05$), although the difference was, as one would expect, in favor of the recognition questions.

Objective Use Questionnaire Data

Based upon their responses to the questionnaire, students in the behavioral objective group were classified as users or nonusers of the objectives. Five nonusers ($M = 18.60$) and 23 users ($M = 25.17$) were identified. When the data were analyzed with the nonusers removed, exactly the same pattern of results emerged as when they were included.

Time Data

When the total amount of study time as measured in minutes was compared for the behavioral and nonbehavioral objective groups, students given the behavioral objectives ($M = 18.39$, $SD = 4.51$) spent significantly more time studying ($t = 2.85$,

[2] This procedure was suggested by Gene V. Glass, personal communication, January 20, 1971.

$df = 54$, $p < .01$) than students given the nonbehavioral objective ($M = 15.29$, $SD = 3.59$). However, when the amount of time the students spent reading the directions before reading the passages was subtracted from the total study time, the difference between the behavioral objective ($M = 16.04$, $SD = 4.23$) and the nonbehavioral objective groups ($M = 14.57$, $SD = 3.40$) was nonsignificant ($t = 1.42$, $df = 54$, $p > .05$).

Discussion

It was thought that having behavioral objectives during study would be helpful to the student on questions which his past history did not lead him to expect. Because most teacher-made tests are composed of items requiring recall of information, it was predicted that behavioral objectives would prove helpful on application questions but not on recognition items. Just the opposite was found in the experiment. Two aspects of the data led to the speculation that the importance of an idea, as judged by the students based upon the nature of the textual materials, was responsible for the results.

The data showed that if the student could correctly identify the definition of a term, he was also relatively successful at answering the related application questions (63% of the time all three of the related application questions were correctly answered; 22%, two of the three; 10%, one of the three; and 5%, none of the three). If, however, the student was unable to correctly define the term, he was relatively unsuccessful at answering the related application questions (15% of the time all three of the related application questions were correctly answered; 19%, two of the three; 25%, one of the three; and 41%, none of the three). This trend held for both the behavioral and nonbehavioral objective groups. This suggests that the classification of an item of information as important enough to be tested was the important dimension rather than the anticipation of a type (recognition or application) of question. It should be noted that since this breakdown of the data involves only one third of the recognition questions, this relationship does not suggest that identical findings should result from all recognition and application questions. Similar results should be found when the recognition questions test the same items as the application questions or when both are classified as important enough to be tested, regardless of specific content.

Further support for the notion that the judged importance of an item of information was the determining factor came from an inspection of the recognition data which indicated that the difference between the behavioral and nonbehavioral objective groups was created primarily by the one question from each passage requiring a psychologist's name as an answer. The difference between the behavioral objective and nonbehavioral objective means of 2.00 raw score points when all recognition questions were considered was reduced to .32 raw score points when the name questions were removed. It appears that students given the nonbehavioral objective anticipated the definition questions but did not anticipate the name questions. Contextual cues may explain this. Because the study passage was entitled "Prompting" not "People Who Have Studied Prompting" and the majority of the text dealt with what prompting is and its two different types instead of telling about the psychologists, the student may have decided that his task was to understand prompting. Knowing that "Anderson" is somehow connected with prompting has very little to do with whether the process of prompting is understood; therefore, the student may have classified this item of information as unimportant and hence unlikely to be tested.

Experiment 2

A second study was designed to test the hypothesis that the judged importance of an item of information determines whether knowledge of the behavioral objectives during training is helpful. Providing students with behavioral objectives during study should improve posttest performance on items students judge unimportant (not measuring how well they know and understand the tested material) but should not influence posttest performance on items students

judge important. A delayed posttest was included to discover whether these predicted differences would persist.

Method

Subjects

College seniors in two sections of an educational psychology course were randomly assigned to the two experimental conditions so that approximately half of each section was in each condition. Of the 32 subjects originally assigned to the behavioral objective condition, 2 were absent the day of the delayed posttest, leaving 30 subjects in the behavioral objective group. Of the 34 subjects originally assigned to the nonbehavioral objective group, 3 were absent the day the posttest was administered and 1 was randomly withdrawn to maintain equal *n*s, leaving 30 in the group. An additional group of 27 students in another section of the course was used as the control group.

Materials

Passages. Five passages were used.[3] Four were modified versions of those used in the first study. The fifth was on overlearning and replaced the passage on test–retest reliability which had a higher mean correct than the other four passages in the first study.

This time no underlining was used and two dates were added to the second paragraph of each passage. The dates were chosen so that each appeared only once. The passages remained approximately 410 words in length.

Posttest. The test consisted of four-alternative multiple-choice items which were randomly ordered. The letter of the correct choice was also randomly ordered.

As in the previous study, the test consisted of both recognition and application questions. However, this time two types of recognition items were used.

One set of three recognition items were prepared for each passage that would probably be classified as "not measuring how well I grasped the material" (unimportant). These consisted of two questions requiring psychologist's names and a third item requiring one of the two dates as answers. The tested dates were chosen so that for three of the passages the date presented first in the reading was tested, while the second date was tested in the remaining two passages.

A second set of three recognition items for each passage required the definitions of the three terms introduced by the passage. Students would probably classify these as "measuring how well I grasped the material" (important).

The remaining questions were three application questions from each reading passage. These questions were in the same form as the application questions in the first study; all dealt with the major process introduced by the passage, and all included descriptions of situations different from those described in the reading passage.

Where possible, test items from the previous study were used for the second study. Items were revised to make the incorrect alternatives more plausible when item analysis suggested such revisions.

Objectives. The behavioral objectives were the same as in the first study with one modification suggested by the students to make them easier to understand. Each objective began with the phrase, "On the test you will be given" and continued as before.

The nonbehavioral objective was the same as in the first experiment except now students were told the test was a 45-item test.

Objective use questionnaire. The same questionnaire was used as in the first experiment to identify students who did not use the behavioral objectives.

Importance questionnaire. The importance questionnaire had the student examine each test item and classify it either as "does" or "does not show how well you grasped the material in the reading passages." This let the students judge and classify the tested items of information as to their importance.

Procedure

The procedure was that used in the first experiment with the following changes. The directions told the students they were participating in an experiment "designed to investigate how people learn from written prose materials," since students in the first experiment were convinced they were in an experiment.

A third manila envelope was given each student in the experimental groups after they completed the posttest. This envelope contained the importance questionnaire and a second copy of the test.

In this experiment, the time students copied from the blackboard was changed every 10 seconds rather than every minute.

When the students returned to class 11 days later following a period of observation in the public schools, the students in the experimental groups without warning or without feedback from the first test were again given the posttest and were asked to record the time when they began and finished the exam.

---

[3] The passages, test, and questionnaires have been deposited with the National Auxiliary Publications Service. Order NAPS Document No. 02174 from ASIS/NAPS, % Microfiche Publications, 305 East 46th Street, New York, New York 10017. Remit with order for each NAPS document number $1.50 for microfiche or $5.90 for photocopies. Make checks payable to Microfiche Publications.

TABLE 2
MEANS AND STANDARD DEVIATIONS FOR THE POSTTEST SCORES IN EXPERIMENT 2

Group/item type	Posttest		Delayed posttest	
	M	*SD*	*M*	*SD*
Behavioral				
Unimportant recognition	9.97	3.81	6.30	2.22
Important recognition	12.60	1.96	11.66	1.96
Important application	11.83	2.32	12.40	2.03
Nonbehavioral				
Unimportant recognition	6.17	3.33	4.30	2.28
Important recognition	12.50	1.81	11.57	2.31
Important application	11.87	2.91	11.83	2.76

Results

Importance Questionnaire Data

As anticipated, a definite pattern emerged from the importance questionnaire data. Application items were judged important by most students. The percentage of students judging single application items as important ranged from 68% to 98% (*Mdn* = 89.25%). Recognition items requiring definitions were also judged important, with individual item agreement ranging from 80% to 98% (*Mdn* = 94.38%). On the other hand, as anticipated, recognition questions requiring dates and names were classified as unimportant. The percentage of students judging single-name items as unimportant ranged from 72% to 82% (*Mdn* = 77.17%), while the date questions ranged from 89% to 91% (*Mdn* = 89.13%).

Based upon the importance questionnaire data, the posttest questions were subdivided into three types of items: one third were unimportant recognition items (including both name and date questions), one third were important recognition items (the definition items), and the remaining third were important application items.

Test Data

The means and standard deviations for the experimental groups subdivided by the three item types are given in Table 2. A two-tailed comparison of the control group ($M = 19.59$, $SD = 3.18$) and the two experimental groups combined ($M = 32.47$, $SD = 6.65$) indicated that the experimental groups performed significantly better ($t = 12.26$, $df = 85$, $p < .01$), supporting the notion that learning had occurred. The means of the control group on the individual item types (unimportant recognition, 3.40; important recognition, 8.07; and important application, 8.03) show improvement occurred on all item types.

Nine two-tailed orthogonal comparisons were made using the same procedure outlined in the first experiment. As predicted on the posttest, the behavioral objective group performed significantly better than the nonbehavioral objective group on the unimportant recognition questions ($t = 6.91$, $df = 319$, $p < .01$), while the differences between these two treatment groups on the important recognition questions ($t = .30$, $df = 319$, $p > .05$) and the important application questions ($t = .06$, $df = 319$, $p > .05$) were nonsignificant.

On the delayed posttest, the same pattern of results held in that the differences between the behavioral and nonbehavioral groups on the important recognition questions ($t = .18$, $df = 319$, $p > .05$) and the important application questions ($t = 1.03$, $df = 319$, $p > .05$) were nonsignificant, while the behavioral objective group performed significantly better than the nonbehavioral objective group on the unimportant recognition questions ($t = 3.64$, $df = 319$, $p < .01$).

Posttest–delayed-posttest comparisons revealed significant losses for the behavioral and nonbehavioral groups combined on the unimportant recognition questions ($t = 7.11$, $df = 319$, $p < .01$) and the important recognition questions ($t = 2.40$, $df = 319$, $p < .05$), but no significant loss on the important application questions ($t = .69$, $df = 319$, $p > .05$) where a slight gain was shown due to the behavioral objective group.

Objective Use Questionnaire Data

Only 1 subject of the 30 in the behavioral objective group reported that he had not used the objectives. Two subjects failed to

fill in the questionnaire. The data analyzed with these subjects removed resulted in the same pattern of results as are reported for the complete sample.

Time Data

When total study time on the passages was compared, the behavioral objective group ($M = 25.60$, $SD = 8.31$) spent significantly more minutes on the task ($t = 2.88$, $df = 58$, $p < .01$) than the nonbehavioral objective group ($M = 19.66$, $SD = 7.60$). However, when the length of time spent initially reading the directions was subtracted from the total study time, the difference between the behavioral ($M = 22.74$, $SD = 8.25$) and the nonbehavioral ($M = 18.79$, $SD = 7.55$) groups was nonsignificant ($t = 1.94$, $df = 58$, $p > .05$). The differences between the treatments in the time spent taking the posttest ($t = .03$, $df = 58$, $p > .05$) and the delayed posttest ($t = .29$, $df = 58$, $p > .05$) were both nonsignificant.

Discussion

Much of school learning involves learning from meaningful prose. Therefore, it is important to find ways to increase the probability that students will process the parts of the material relevant to the task set for them. Behavioral objectives are one way to guide the learner's attention; however, previous studies did not consistently show learning advantages for students provided objectives. The data of this study support the hypothesis that the judged importance of an item of information determines whether knowledge of behavioral objectives during training is helpful. Students receiving behavioral objectives during study performed significantly better on test questions that the majority of the students classified as unimportant than students given a nonbehavioral objective. This difference appeared in both the posttest and the delayed posttest data indicating that the difference remained 11 days later. Students given behavioral objectives on test questions that the majority of the students classified as important did not perform significantly differently from students given the nonbehavioral objective on either the posttest or the delayed posttest. This was true for both application- and recall-level questions.

Other ways of guiding the learner's attention to specific parts of prose passages may very well operate in a similar manner. Such techniques should follow the hypothesis used in this study for behavioral objectives only if they consistently set off information needed when the learner is tested and if the learner is aware of this relationship. The underlining in the passages used for the first experiment was not accompanied with directions stating that the underlined words were items that would be tested. Not all items that were underlined were tested and some items that were not underlined were tested; therefore such underlining does not meet the conditions described above. At least one study has shown benefits from the experimenter underlining when directions tell students they will be tested over the underlined portions (Crouse & Idstein, 1972).

An important question that remains to be investigated is whether the hypothesis supported by these studies will generalize to other materials. The length of the study passage, the number of ideas presented, and the length of study time are variables that might affect the generalizability of the hypothesis.

As in the first experiment, the time data revealed that the behavioral objective group spent more time on the total task than the nonbehavioral objective group. Since the directions to the behavioral objective group were much longer, the time the students spent initially reading the instructions was subtracted from the total study time. When the adjusted scores were compared, the difference between the two groups was nonsignificant in both studies. Although no difference was found in the length of study time, the groups may have used their time differently. Logically, behavioral objective students might spend some of their time going back and rereading parts of the objectives, while nonbehavioral objective students would have less reason to do this. Some students were observed following this logic; however, there was no way to measure these behaviors objectively. If these patterns of behavior were fairly general, it is possible the behavioral objective group spent less of their

study time on the reading passages than the nonbehavioral objective group.

As expected, students performed less well 11 days later on the immediate posttest. This finding held for all question types with one exception, the important application questions, in which no significant difference was found. This was not predicted and needs further investigation to see if similar results can be obtained again.

REFERENCES

BLOOM, B. S. (Ed.) *Taxonomy of educational objectives: Handbook 1, cognitive domain.* New York: David McKay, 1956.

CROUSE, J. H., & IDSTEIN, P. Effects of encoding cues on prose learning. *Journal of Educational Psychology*, 1972, 63, 309–313.

HAYS, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

JENKINS, J. R., & DENO, S. L. Influence of knowledge and type of objectives on subject-matter learning. *Journal of Educational Psychology*, 1971, 62, 67–70.

MAGER, R. F. *Preparing instructional objectives.* Palo Alto, Calif.: Fearon, 1962.

MERRILL, P. F., & TOWLE, N. J. The effects of the availability of objectives on performance in a computer-managed graduate course. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

MORSE, J. A., & TILLMAN, M. H. Effects on achievement of possession of behavioral objectives and training concerning their use. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

OSWALD, J. M., & FLETCHER, D. Effects of varying level of specificity and the cognitive level of explicit instructional objectives for use with eleventh grade social science students. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, February 1970.

ROTHKOPF, R. Z. Shopping lists, shopping carts, memory store. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

ROTHKOPF, E. Z., & KAPLAN, R. Exploration of the effect of density and specificity of instructional objectives on learning from text. *Journal of Educational Psychology*, 1972, 63, 295–302.

THORNDIKE, R. L., & HAGEN, E. P. *Measurement and evaluation in psychology and education.* (3rd ed.) New York: Wiley, 1969.

WEBB, A. B., & CORMIER, W. H. Effects of the use of behavioral objectives and criterion evaluation upon the classroom progress of adolescents. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

YELON, S. L., & SCHMIDT, W. H. The effect of objectives and instructions on the learning of a complex cognitive task. Paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

(Received March 12, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 233-237

EFFECTS OF IMAGERY ON LEARNING INCIDENTAL MATERIAL IN THE CLASSROOM[1]

FRANK GOLDBERG[2]

Teachers College, Columbia University

Two groups of 108 fifth-grade children were presented with illustrated or nonillustrated incidental information that served as content material for a spelling and a grammar exercise. Incidental learning was determined by two tests measuring retention of the content material. The incidental-learning test performance of good and poor readers and scores obtained on recognition versus recall questions were also compared. Results indicated that (*a*) illustration facilitated incidental-learning retention, (*b*) good readers retain more incidental learning than poor readers, and (*c*) more recognition than recall questions were answered correctly. The facilitating effect of illustration was explained according to an imagery hypothesis.

The primary purpose of this study was to determine whether imagery has the same enhancing effects in the classroom upon incidental learning as that found in intentional learning. Studies have repeatedly demonstrated that intentional learning, in which specific instructions to learn are given, is a far superior form of learning than incidental learning (Brown, 1954; Jenkins, 1933; Postman, Adams, & Bohm, 1956). The image-evoking ability of words has been shown to facilitate the retention of intentionally learned verbal material. The ease of visual representation of words ranges from abstract nouns, to concrete nouns, to pictures and objects, in increasing order of concreteness (Paivio, 1969). Paivio proposed a two-process theory to explain why imagery actually works. He suggested that stimuli such as concrete words produce both verbal and perceptual codes since their meaning is derived through association with both concrete objects and other words. Abstract terms such as "justice," however, derive their meaning primarily through verbal experiences and are, therefore, likely to arouse only verbal storage codes. The two-process theory thus identifies imagery effects with amount of information stored as a result of its additional visual code.

Imagery, as used in this study, refers to the illustration of learning materials that are incidental to a central learning task. These illustrations may thereby directly create an additional visual representation of the incidental stimuli, which should facilitate incidental-learning retention, according to Paivio's (1969) theory.

The importance of this study lies in its practical application to curriculum design. The implication is that one subject area can be included as ancillary content material for an *entirely different lesson.* The end result is a coordinated curriculum in which the learning of two or more areas is achieved at little additional investment of cost or time.

The relationship between reading ability and incidental learning and a comparison of recall versus recognition scores obtained on incidental-learning tests were also included in the present investigation. The purpose was to broaden the scope of the study while ascertaining whether the effects of illustra-

[1] This article is based on a doctoral dissertation submitted to Teachers College, Columbia University. Members of the dissertation committee included Mary Alice White, chairman, N. Dale Bryant, and John Nolan. Their suggestions and assistance are gratefully acknowledged.

[2] Requests for reprints should be sent to Frank Goldberg who is now with the Office of Special Education and Pupil Personnel Services, Evaluation and Placement Unit, P.S. 206, Neck Road and East 22nd Street, Brooklyn, New York 11229.

tion are comparable among children of different reading levels and on two types of response scores.

Finally, a substudy compared the retention of material learned intentionally or incidentally. Previous research has tended to use word lists or paired-associates as experimental materials. The present study, however, sought to determine whether the oft-found superiority of intentional over incidental learning would also be replicated when task-specific procedures and curricular learning materials are employed. Such parallel results would provide justification for the use of curricular materials in incidental-learning studies.

Method

The research was conducted in two different classroom curricular areas. The incidental-learning material was presented as part of a spelling lesson in Part 1 and as content material of a grammar exercise in Part 2. The rationale for including a second task was to broaden the scope of information employed as incidental material and to compare the results with Part 1, in which different procedures and learning materials were used. Part 2 was completed approximately four to five weeks after Part 1. The subjects were the same for both parts of the study.

Subjects

The sample consisted of 216 fifth-grade children (113 boys, 103 girls). They were the students of 10 classes selected from three public schools located in the same area. The subjects within each class were paired according to reading scores obtained on the Metropolitan Achievement Test (Form F, Elementary Grades) which was administered to them in the fourth grade. Each member of the matched pair was then randomly assigned to the imagery or nonimagery group in an A–B–B–A pattern. The imagery and nonimagery groups differed by the presence or absence of illustrations for their incidental-learning materials. The mean reading level for each group of 108 subjects was 5.25. The effect of presentation method (imagery versus nonimagery) upon incidental-learning retention represented the major variable investigated in this study.

The imagery and nonimagery groups were then further subdivided into poor readers and good readers. The reading groups were obtained by calculating the mean reading score for each of the 108 imagery and nonimagery matched pairs. Pairs whose mean reading level was below the median of 5.2 constituted the poor readers and those above the median, the good readers. The imagery and nonimagery groups were thus evenly divided between 54 poor and 54 good readers. The mean Metropolitan Achievement Test scores were 3.9 and 6.6, respectively, for both groups. Comparison of poor and good readers on incidental-learning retention was the second variable examined by this study.

Materials and Procedure

Part 1. The spelling lesson consisted of 24 sentences, each containing a blank space in which a spelling word was to be written. Each sentence consisted of varying levels of information relating to social studies or science facts that served as the incidental-learning material. Two sets of identical spelling texts were prepared with the imagery group having, in addition, their incidental-learning material illustrated by a picture below each sentence that presented in visual form the information given in the text. For example, the spelling item, "Income (*tax*) must be paid before April 15th in the United States," was illustrated by a picture of an April calendar with the "15" prominently circled.

Uniform instructions were given to each teacher to dictate each spelling word, the accompanying sentence demonstrating its usage, and finally, the word again. The children were then instructed to write each spelling word in its corresponding blank space. Immediately following completion of the spelling lesson, a test was administered to all subjects which measured their retention of the incidental-learning material. The test was composed of 12 recall (fill-in) and 12 recognition (multiple-choice) questions, each one corresponding to information given in 1 of the 24 sentences. The items were counterbalanced for sequence with half of the subjects tested first on the recall questions and half on the recognition questions. One half hour was allotted to complete the incidental-learning test. Comparison of recall and recognition incidental-learning scores constituted the third variable explored by the present investigation.

It should be noted that all incidental-learning material used in this study was first piloted on a separate group of 25 intellectually gifted fifth-graders to rule out any possibility that the information was already known to the subjects.

Part 2. The incidental-learning material for the grammar exercise consisted of a social studies lesson on China and a science lesson on whales, each approximately one page in length. As in Part 1, the imagery and nonimagery texts differed by the presence or absence of illustrations.

The teachers were instructed to review with their classes proper and common nouns and how to recognize them in sentences. The social studies and science lessons were then presented to each class as practice exercises in noun recognition with instructions to circle the proper nouns and underline the common nouns. Forty-five minutes were given to complete the grammar exercises. Immediately following completion of the exercises, a

test measuring retention of the incidental-learning material was administered to all subjects. As in Part 1, the test consisted of 24 questions divided into 12 recall and 12 recognition questions. One half of the questions pertained to the China lesson and the other half to the lesson on whales. The test had a 30-minute time limit.

Certain precautions were taken to minimize the possibility of any subject anticipating the incidental-learning test of Part 2 as a result of his prior experience with Part 1. The measures taken included

1. Four to five weeks had elapsed between Parts 1 and 2.
2. No materials were distributed to the teachers in their students' presence as was done in Part 1. This procedure precluded any chance of the investigator being seen as an associative link between both parts of the study.
3. A postexperimental inquiry was conducted by each teacher to determine whether any subject anticipated the incidental-learning test or deliberately rehearsed the material (Postman, 1964). Four subjects were eliminated from the study as a result of the inquiry.

Statistical Analysis

The research design was in the form of a 2 × 2 × 2 analysis of variance with repeated measures on the first two factors (Winer, 1962). The three factors were Presentation Method (Imagery versus Nonimagery) × Type of Question (Recall versus Recognition) × Reading Level (Poor versus Good Readers). Parallel but separate analyses were performed for both parts of the study.

Substudy

Subjects and Procedure

An additional fifth-grade class comprised of 18 subjects from a fourth public school constituted the intentional-learning group. The entire class was given the nonillustrated grammar exercises of Part 2 with the additional instructions to learn the information while circling or underlining the nouns. Immediately thereafter, the incidental-learning test was administered. All time allotments were the same as in Part 2. The incidental-learning test scores of the intentional-learning group were compared to the incidental-learning group (nonimagery group, Part 2) in order to determine the amount of retention occurring in the presence or absence of instructions to learn.

Statistical Analysis

An analysis of covariance, with reading achievement as the covariate, was performed since the intentional- and incidental-learning groups were not prematched according to reading ability as in the main study.

Results

The results for both parts of the main study are presented and analyzed together since the findings were so consistently similar.

Table 1 presents the means and standard deviations for all three variables. The higher incidental-learning scores obtained by the imagery group in both curricular areas clearly demonstrates that illustrated incidental material was better retained than nonillustrated material. The means for the good readers were also greater than those for the poor readers in both subject areas. Finally, the subjects scored higher on incidental-learning recognition questions than on recall questions.

An analysis of variance indicated that all main effects were highly significant while none of the interactions were significant. The F ratios for presentation method were 9.76 ($df = 1/106$, $p < .005$) for Part 1 and 16.25 ($df = 1/106$, $p < .0005$) for Part 2. The comparable F values for reading level were 107.75 ($df = 1/106$, $p < .0005$) and 118.94 ($df = 1/106$, $p < .0005$). Finally, the F ratios for type of question were 50.19 for Part 1 ($df = 1/106$, $p < .0005$) and 43.08 for Part 2 ($df = 1/106$, $p < .0005$).

TABLE 1

Mean Number of Correct Responses and Standard Deviations for Incidential Learning Tests of Main Study

Curricular areas	Presentation method				Reading level				Type of question			
	Imagery	*SD*	Non-imagery	*SD*	Good	*SD*	Poor	*SD*	Recog-nition	*SD*	Recall	*SD*
Part 1, spelling	12.57	5.35	11.05	5.73	15.30	4.69	8.30	4.00	6.52	3.49	5.27	2.95
Part 2, grammar	13.36	5.39	11.27	5.50	15.79	4.17	8.82	4.45	6.83	3.17	5.48	3.05

Thus, each of the three variables asserts its effect independently across both levels of the other two factors.

An analysis of covariance for the substudy yielded adjusted incidental-learning means of 15.27 for the intentional-learning group and 11.48 for the incidental-learning group. The difference was significant ($F = 12.29$, $df = 1/123$, $p < .0005$), indicating that material which is learned intentionally is better retained than material which is learned incidentally.

Discussion

The findings indicate that illustration can facilitate the retention of incidental material in the classroom regardless of reading ability or method of questioning. Furthermore, the striking similarity of results obtained in Parts 1 and 2, despite the use of entirely different procedures and subject matter, serves to lend further credence to the findings and to their generalization beyond the confines of the present study.

The superiority of the good over the poor readers in incidental-learning retention apparently indicates that, as in intentional learning, the brighter students and better readers are usually the better incidental learners. The finding that more recognition than recall questions were answered correctly is also in agreement with the results of previous studies.

The investigator attributed the effects of illustration to imagery. Alternative theoretical explanations should, however, also be considered. For example, the superiority of illustrations can also be attributed to their ability to attract and maintain the subject's attention to the incidental material. The pictures, in effect, could serve to focus one's attention on the incidental-learning facts, resulting in the subject spending more effective time on the relevant information in the text (Simon & Jackson, 1968). A second explanation is that illustrations may serve as attention-capturing devices by arousing curiosity (Paradowski, 1967).

The following postexperimental analysis was performed in order to clarify the theoretical implications. It was reasoned that the relative amount of illustration is important according to an imagery hypothesis, whereas it is not crucial for either of the two "attentional" hypotheses. For example, one of the incidental items of Part 1 consisted of the statement, "A (*frog*) has four toes on his front legs and five toes on his hind legs." The illustration accompanying this sentence showed a detailed drawing of a frog with four front toes and five hind toes. For an imagery interpretation, the visual duplication of information provided in the text is a must. The reason for this is that the two-process theory attributes the effects of imagery to the amount of information stored as a result of its additional visual code. Consequently, only a detailed illustration paralleling the information provided by the text is helpful in answering the subsequent incidental-learning test question, "How many toes does a frog have on his front legs?" According to an attentional interpretation, however, a picture of only a frog's head might have served the purpose equally well since the illustration serves only an attentional, rather than an information-providing function.

The analysis was done in the following manner. Two psychology graduate students were asked to rate each of the illustrated spelling and grammer facts as to whether they were high or low on imagery. Guidelines for receiving a high rating included (*a*) clarity of the drawing and (*b*) how close it came to visually replicating the incidental-learning information provided in the text. A third graduate student was used for the seven spelling and eight grammer items on which the first two judges disagreed.

It was then hypothesized that, according to an imagery interpretation, a greater percentage of high-imagery items should be passed by the imagery group than those items rated low on imagery. No such difference was predicted according to the attentional hypotheses.

Results showed that 57.5% of the high-imagery items were passed compared to 57% for the low-imagery items. Comparative results for grammar were 63.2% and 54.1%, respectively ($\chi^2 = 7.4$, $df = 1$, $p < .01$). The grammar findings thus suggest an

imagery interpretation, whereas the spelling results are more in agreement with an attentional explanation.

Of particular pertinence to the theoretical argument is a descriptive occurrence during the study that cannot be quantified. The investigator, while grading the incidental-learning tests, came across 15 papers in which the subjects actually drew their own illustrations near the correct answer, presumably as a mnemonic device. Even more dramatic were the three subjects who drew pictures in the answer spaces with explanatory notes to the effect that they could not recall the answers but knew what they looked like. These test behavior phenomena point strongly to an imagery interpretation and support Bahrick's (1969) findings that verbal and visual components of a memory trace can be forgotten independently.

The practical educational implications of this study are numerous and varied and have direct application to the new instructional technology recommended for the classroom (Carnegie Commission on Higher Education, 1972). For example, instead of including relatively unimportant information as ancillary material for spelling lessons or grammar exercises, the present study suggests that it would be far more worthwhile to employ educationally informative material. Furthermore, it is preferable to include material that is being presented for the first time or that needs review rather than material with which the students are already familiar. Illustration would then be a means of assuring greater incidental-learning retention as demonstrated by the results of this investigation.

REFERENCES

Bahrick, H. P. Discriminative and associative aspects of pictorial paired associate learning: Acquisition and retention. *Journal of Experimental Psychology*, 1969, **80**, 113–119.

Brown, G. H. Factors influencing incidental learning. *Journal of Experimental Psychology*, 1954, **47**, 163–169.

Carnegie Commission on Higher Education. *The fourth revolution*. Hightstown, N.J.: McGraw-Hill, 1972.

Jenkins, J. R. Instruction as a factor in incidental learning. *American Journal of Psychology*, 1933, **45**, 471–477.

Paivio, A. Mental imagery in associative learning and memory. *Psychological Review*, 1969, **76**, 241–263.

Paradowski, W. Effect of curiosity on incidental learning. *Journal of Educational Psychology*, 1967, **58**, 50–55.

Postman, L. Short term memory and incidental learning. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press, 1964.

Postman, L., Adams, P. A., & Bohm, A. M. Studies in incidental learning: V. Recall for order and associative clustering. *Journal of Experimental Psycholgy*, 1956, **51**, 334–342.

Simon, S. H., & Jackson, B. Effect of a relevant versus irrelevant observation stimulus on concept-identification learning. *Journal of Experimental Psychology*, 1968, **76**, 125–128.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received March 13, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 238-246

MORNING-TO-AFTERNOON CHANGES IN COGNITIVE PERFORMANCES AND IN THE ELECTROENCEPHALOGRAM[1]

EDMUND J. MACKENBERG[2]

Clark University

DONALD M. BROVERMAN[3] AND WILLIAM VOGEL

Worcester State Hospital and Clark University

EDWARD L. KLAIBER

Worcester State Hospital and Worcester Foundation for Experimental Biology

Sixty male subjects were tested, both in the a.m. and p.m., in a Latin square design, on six cognitive tests and on the electroencephalogram (EEG). Performances of overlearned repetitive tasks were found to be better in the morning than in the afternoon, while the reverse pattern of performance was found for perceptual-restructuring tasks. The EEG was found to be more sensitive to photic stimulation in the afternoon than in the morning. The findings were interpreted as reflecting a hormonally mediated shift from adrenergic to cholinergic dominance in the brain, from morning to afternoon. Pedagogical implications are discussed.

This study tests the hypothesis that performances of overlearned, serially repetitive *automatized* tasks and perceptual-restructuring tasks change in opposite ways from morning to afternoon; that is, performances of overlearned repetitive tasks are expected to be better in the morning than in the afternoon and performances of perceptual-restructuring tasks are expected to vary in the reverse manner. *Perceptual-restructuring* tasks are defined as novel tasks in which perceptually obvious stimulus attributes do not facilitate task solution but, rather, act to conceal or obscure the perception of nonobvious but correct stimulus attributes, for example, the Embedded Figures Test (Witkin, 1950).

The psychological and physiological processes underlying performances of serially repetitive automatized tasks, for example, speed of color naming or simple additions, and perceptual-restructuring tasks have been the object of theoretical concern (Broverman, 1964; Klaiber, Broverman, & Kobayashi, 1967). At a psychological level of discourse, it has been hypothesized that variations in performances of overlearned repetitive tasks are a positive function of the strength of response to appropriate, perceptually obvious stimulus cues.

On the other hand, performances of perceptual-restructuring tasks are thought to be, in part, a negative function of the strength of response to obvious stimulus attributes that, in this case, are not appropriate since they tend to obscure the correct, nonobvious stimulus attributes. Thus, negative intrapsychic relationships are postulated to exist between the psychological processes underlying performances of overlearned, simple repetitive tasks versus performances of perceptual-restructuring tasks. First-order correlations between such tasks, however, are typically positive, though small (for example, Podell and

[1] Based on a dissertation by the senior author submitted in partial fulfillment of the requirements for the doctoral degree at Clark University.

[2] Now with the Department of Corrections, Deuel Vocational Institution, Box 400, Tracy, California 01910.

[3] Requests for reprints should be sent to Donald M. Broverman, Department of Psychology, Worcester State Hospital, Worcester, Massachusetts 01613.

Phillips, 1959). The reason for the positive, rather than the expected negative, correlations has been postulated to be due to the overriding effect of a general ability factor present in both categories of tasks (Broverman & Klaiber, 1969). Demonstration of the postulated negative relationship may, therefore, require assessments of both tasks in the same individuals under varying conditions, when the conditions are known to affect at least one of the two categories of tasks. In this manner, the general level of performance of an individual is held constant, while conditions known to affect one category of task, for example, overlearned repetitive tasks, are varied. The opposite pattern of change in performance may then be predicted for the remaining category of task, in this case, perceptual-restructuring tasks; that is, if a given condition is known to adversely affect performances of automatized tasks, the same condition should enhance performances of perceptual-restructuring tasks.

Performances of such serially repetitive tasks as speed of color naming and simple addition problems have been reported to be better in the morning than in the afternoon (Hollingworth, 1914; Muscio, 1920). The decline in performance from morning to afternoon was attributed to cumulative fatigue arising from normal daily activities (Hollingworth, 1914).

Therefore, since performances of simple repetitive tasks are known to be better in the morning than in the afternoon, the present study predicts that performances of perceptual-restructuring tasks will vary in the opposite manner; that is, perceptual-restructuring tasks should be performed better in the afternoon than in the morning.

Physiologically, it has been suggested that performances of simple repetitive tasks are dependent upon central adrenergic neural processes, while performances of perceptual-restructuring tasks are dependent upon central cholinergic neural processes (Broverman, Klaiber, Kobayashi, & Vogel, 1968). This hypothesis stems from reports that adrenergic stimulants tend to enhance performances of overlearned repetitive tasks and to impair performances of perceptual-restructuring tasks, while cholinergic stimulants have the reverse pattern of effects (Broverman et al., 1968). Accordingly, indices of central adrenergic functioning should be higher in the a.m. than in the p.m. The present study tests this hypothesis with an electroencephalogram (EEG) index, EEG "driving" in response to photic stimulation. EEG *driving* refers to the tendency of the EEG waves to synchronize with a rapidly flashing bright light placed before the closed eyes of a subject. Adrenergic stimulants tend to diminish this effect, while adrenergic blocking agents tend to enhance the phenomenon (Vogel, Broverman, Klaiber, & Kun, 1969). Therefore, the present study predicts that significantly less EEG driving will occur in the morning than in the afternoon.

Method

Subjects

Sixty paid, normal male volunteer subjects participated in the study. Because the subjects ranged widely in age (from 12 to 50 years), they were trichotomized into three equal-sized age groups to check for possible age differences in morning-to-afternoon changes in performance. The age groupings were as follows, with 20 persons in each group: (*a*) from 12 to 18 years ($M = 14.0$, $SD = 2.12$), (*b*) from 19 to 23 years ($M = 20.4$, $SD = 1.53$), and (*c*) from 24 to 50 years ($M = 28.7$, $SD = 2.74$).

Subjects with known physical problems were excluded, as were all subjects taking any sort of medication or who admitted to taking drugs.

Cognitive Tasks

Overlearned Repetitive Tasks

Speed of color naming (Broverman, 1964). This task requires the subject to name, as fast as possible, 10 lines of randomly arranged red, green, and blue color patches, 10 patches to the line. The time taken to complete the task is taken as the score for the test.

Speed of naming repeated objects. Subjects are asked to name, as fast as possible, 100 randomly arranged pictures of a cup, a tree, and a fly, 10 pictures to the line. The time taken to complete the task is taken as the score.

Errors on simple, serially repetitive tasks such as these are rare. Scores corrected for errors have been found to correlate .9+ with uncorrected scores (Stroop, 1935). Hence, errors are typically not recorded.

Perceptual-Restructuring Tasks

Embedded Figures Test (*Witkin, 1950*). This task requires the subject to find a simple geometric figure embedded within a complex pattern of lines and colors. To make two alternative sets, 5 simple figures (A, C, D, E, and G) and 10 complex figures were used. The two sets were as follows: Set 1, Figures A1, C2, D2, E1, G2; and Set 2, Figures A3, C3, D1, E5, G1.

The total time taken to identify each of the five figures in a given set was taken as the score for the task. A maximum of 90 seconds was allowed for each figure.

Wechsler Adult Intelligence Scale (*WAIS*) *Block Design subtest* (*Wechsler, 1955*). This task, which requires subjects to reduce the obvious perceptual pattern to parts corresponding to blocks in order to construct the given design, has been reported to load on the same cognitive factor as the Embedded Figures Test (Broverman, 1964). Two sets of block designs were formed: Set 1, the odd-numbered designs; and Set 2, the even-numbered designs.

Choice-of-Response Tasks

Two additional tasks that permit subjects to respond to either obvious or nonobvious embedded stimulus attributes were also included as described below.

Rod-and-frame test (*Witkin, Lewis, Hertzman, Machover, Meissner, & Wapner, 1954*). This test requires the subject to adjust a luminescent rod to the vertical in a darkened room when the rod is within a tilted luminescent square frame. "For successful performance of this task the subject must 'extract' the rod from the tilted frame through reference to body position [Witkin, et al., 1954, p. 25]." The subject, then, must inhibit the influence of the frame.

The data from the rod-and frame situation was scored in a manner consistent with Witkin et al., (1954) in measuring field dependence and field independence. This measure consists of the sum of the absolute deviations from true vertical. In the terms of this study, field dependence would be analogous to heightened responsivity to obvious visual stimuli, that is, greater response to the tilted frame. Field independence would be analogous to a perceptual restructuring of these visual stimuli, that is, evidence of an ability to separate judgments of verticality from the influence of the tilted frame.

Rorschach Test. This test was scored for functional–integrative responses (Phillips, Kaden, & Waldman, 1959) which are said to measure the subject's ability to respond to inkblot stimuli in an inhibitory, restructuring manner. The functional–integrative response is scored when the percept indicates that an interaction between two or more conceptually independent subunits, that is, movement of at least one object in relationship to at least one other object, is occurring. In addition, the percept must contain M (human movement), FM (animal movement), or m (inanimate movement) as a determinant.

Functional–integrative responses were given weighted scores according to the criteria set forth by Phillips et al. (1959). When the movement is active, a weighted score of 4 is given to a functional–integrative response containing active human content and a weighted score of 2 is given to a functional–integrative response with animal content. When the movement is passive or static, a weighted score of 1 is given to functional–integrative responses with either human or animal content. A weighted score of 1 is given to functional–integrative responses with inanimate (neither animal nor human) content.

The Rorschach Test was administered to the subjects in two sets of five cards each, one at each of two experimental sessions. The cards were separated into sets of aproximately equal stimulus value, and average number and quality of responses, according to Phillips and Smith (1953), as follows: Set 1 = Cards, 1, 3, 4, 6, 9, and Set 2 = Cards, 2, 5, 7, 8, 10.

EEG Photic Driving

An EEG, with photic stimulation, was obtained from each subject at both of the experimental sessions. The EEG-driving measure was obtained on a Grass model 5D polygraph, with Grass model 5P5E preamplifiers. The time constant was 2 second and the sensitivity used was 30 microvolts per centimeter pen deflection. Grass EC2 electrode cream was applied under Grass E5 gold disc electrodes. The EEG data obtained results from one pair of occipital electrodes with each electrode placed 3 centimeters to either side of the midline, between the O_1–O_2 and P_3–P_4 derivations.

Recordings were obtained at rest for a period of 5 minutes and during photic stimulation. Stimulation was performed while the subject kept his eyes closed, using a Grass PS2 photo stimulator which was placed parallel to the plane of the face, 30 centimeters from the eyes, so that both visual fields were equally stimulated. Stimulation began at Intensity 2 for 10 seconds, for trials of 5, 10, 15, 20, 25, and 30 flashes per second, with 20 seconds rest between trials. Similar series of trials were given at Intensities 4 and 8. Thus, each subject was exposed to 18 photic stimulation trials.

An EEG-driving response to photic stimulation was defined as two consecutive seconds of EEG waves at the fundamental frequency of the photic stimulation or at one of its harmonics. "Driving" was scored only if the entire response, as defined, occurred within the limits of the duration of the photic signal and only if activity from one frequency band, and no other, was present during a two-second period, that is, superimposed activity was not scored.

Procedure

Equal numbers of subjects in each age group were randomly assigned to Groups A and B. Group

A began the experimental procedure at 9:30 a.m. and was tested again at 3:30 p.m., while Group B began the experimental procedure at 3:30 p.m. and was tested again the following morning at 9:30 a.m. Subjects were requested to avoid all stimulants (coffee, coke, alcohol, etc.) throughout the day and were screened for a good previous night's rest.

Task Administration

The two groups were subdivided into subgroups (A1 and A2, B1 and B2) so that the order of presentation of alternate sets of various tasks (the Embedded Figures, the WAIS Block Designs, and the Rorschach Tests) could be counterbalanced.

This procedure enabled variance due to differences in stimuli to be extracted.

The order of presentation of the tasks to each subject in each of the sessions was as follows:

1. EEG measure,
2. rod-and-frame test,
3. naming color hues test,
4. Embedded Figures Test,
5. speed of naming repeated objects test,
6. WAIS Block Design subtest, and
7. Rorschach Test.

All tasks were administered in the same sequence to each subject in each experimental session in order to minimize possible effects due to order of administration.

Results

Method of Data Analysis

A mixed analysis of variance and Latin square method of analysis was employed. Because the design employs repeated measures for each subject, the between-individuals source of variance was first extracted. Sequence effects related either to a.m. versus p.m., or Set 1 versus Set 2 are included in the between-individuals variance. Since these sequence effects were not of theoretical interest, no further isolation of these effects was carried out. Variance due to age, which is also a part of between-individuals variance, was isolated and evaluated.

The within-individuals variance was analyzed for the following main effects: morning versus afternoon; order (1st administration versus 2nd administration, or practice); and stimuli sets, where appropriate. Interactions of order by morning–afternoon and order by stimuli sets are confounded with sequences and, accordingly, are not interpretable; therefore, they were not computed.

The interactions of age by morning–afternoon, age by order, and age by stimulus sets are not confounded and were, therefore, extracted.

Square uniqueness is confounded with between-individuals variance and, therefore, was not extracted.

Table 1 presents the combined results of the study in summary form. Five of the six experimental hypotheses involving differences in cognitive performance in the morning versus the afternoon received statistically significant support (only the differences in WAIS Block Design performances failed to reach significance, but the means were in the expected direction). The phys-

TABLE 1

Summary of Results

Task and measure	Hypothesis	p[a]	Status
Overlearned repetitive			
Color naming	Subjects faster in a.m. than in p.m.	$<.05$	Supported
Repeated objects	Subjects faster in a.m. than in p.m.	$<.025$	Supported
Perceptual-restructuring			
Embedded Figures	Subjects faster in p.m. than in a.m.	$<.01$	Supported
WAIS Block Designs	Subjects faster in p.m. than in a.m.	*ns*	Not supported
Choice of response			
Rorschach	More field independent scores in p.m. than in a.m.	$<.03$	Supported
Rod and frame	Subjects more field independent in p.m. than in a.m.	$<.06$	Supported
EEG	More EEG-driving responses in p.m. than in a.m.	$<.05$	Supported

[a] Two-tailed tests used throughout.

iological EEG measure also changed significantly in the expected direction.

The specific results are described in detail below.

Table 2 presents summaries of the analyses of variance of the dependent variables, that is, six cognitive tasks and the EEG photic-driving index, and Table 3 presents the mean performance levels of the dependent variables in the a.m. and in the p.m. The significant results obtained from the various sources of variance are discussed below.

Age. Age was significantly related to performances of both simple repetitive tasks, the Embedded Figures Test, and the EEG index. The youngest group appears to have performed markedly less well than the two older groups on these three cognitive tasks and to have had significantly more photic-driving responses than the two older groups.

Time of day. This variable, which is the critical variable in the study, was significantly related to the EEG index and to all of the cognitive tasks, except the WAIS Block Designs, in the expected directions. Thus, as expected from previous studies (Hollingworth, 1914; Muscio, 1920) both simple repetitive tasks were performed better in the a.m. than in the p.m. The performances of both perceptual-restructuring tasks, however, change from the a.m. and the p.m. in the opposite manner; that is, they were better in the p.m. than in the a.m., significantly in the case of the Embedded Figures Test and not with statistical significance in the case of the WAIS Block Designs.

The choice of response tasks also shows the predicted significant differences in morning versus afternoon; that is, the rod-and-frame test indicated greater field independence in the p.m. than in the a.m. ($p < .06$), while the Rorschach shows a greater production of functional–integration responses in the p.m. than in the a.m. Functional–integration responses require that initial inkblot percepts be reorganized into complex, nonobvious percepts.

Finally, more EEG photic-driving responses occurred in the p.m. than in the a.m., indicating a probable shift towards cholinergic dominance in the p.m. compared to the a.m.

Order. The second task administration was significantly better than the first on both simple repetitive tasks and both perceptual-restructuring tasks. The two choice-of-response tasks and the EEG did not show significant order effects.

Stimuli sets. These were significant only for the WAIS Block Designs.

Interactions. Only one noninterpretable interaction, Age × Order on the EEG, achieved statistical significance. The main implication of these negative interactive results is that the various age groups do not appear to differ significantly in their shift in performance from a.m. to p.m.

Discussion

The results of the study provide broad support to the experimental hypotheses; that is, all seven changes were in the predicted direction and six were statistically significant. Thus, the results tend to support the postulated intrapsychic negative relationship between the processes involved in performances of automatized and perceptual-restructuring tasks.

Similarly, the observed changes in cognitive performances support the notion that central adrenergic dominance exists in the morning compared to the afternoon. Thus, performances of automatized tasks, which are known to be enhanced by adrenergic stimulants and impaired by adrenergic blocking agents, were better in the morning than in the afternoon. Similarly, performances of perceptual-restructuring tasks, which are facilitated by cholinergic stimulants and impaired by cholinergic blocking agents, were better in the p.m. than the a.m.

The a.m.-to-p.m. changes in number of EEG responses to photic stimulation also support the notion of greater central adrenergic dominance in the morning than in the afternoon.

Hollingworth (1914) suggested that the morning-to-afternoon deterioration in performance of automatized tasks was due to the cumulative fatigue of normal activity. However, fatigue is a difficult concept to define (Bartley, 1957; Floyd & Welford,

TABLE 2

ANALYSES OF VARIANCE OF COGNITIVE TASKS

Source of variance	*df*	Task: Overlearned repetitive — Color naming *MS*	Color naming *F*	Repeated objects *MS*	Repeated objects *F*	Perceptual restructuring — Embedded Figures *MS*	Embedded Figures *F*	WAIS Block Designs *MS*	WAIS Block Designs *F*	Choice of response — Rod and frame *MS*	Rod and frame *F*	Rorschach *MS*	Rorschach *F*	EEG — Photic driving *MS*	Photic driving *F*
Between individuals	59	174.1		179.4		12,854.9		22.9		831.0		20.5		47.2	
Age (A)	2	1150.6	7.96***	905.3	6.04***	45,850.5	3.92**	44.1	*ns*	2407.8	*ns*	6.1	*ns*	139.2	3.22**
Pooled individuals (B)	57	144.4		149.9		11,696.1		21.9		775.7		21.0		43.1	
Within individuals	60	16.0		25.2		3,820.9		5.8		192.5		6.9		63.6	
Time of day (C)	1	44.4	4.04**	80.0	5.60**	20,228.7	7.93***	6.1	*ns*	623.4	3.67*	33.1	5.29**	36.3	7.49***
Order (D)	1	226.9	20.66	580.8	40.7****	51,584.0	20.23****	31.0	11.07****	674.5	3.97*	.7	*ns*	3.0	*ns*
Stimuli sets (E)[a]	1	—	—	—	—	1,903.0	*ns*	57.4	12.26****	—	—	.1	*ns*	—	—
A × C	2	26.1	*ns*	14.2	*ns*	4,437.6	*ns*	1.3	*ns*	137.0	*ns*	.4	*ns*	9.1	*ns*
A × D	2	28.0	*ns*	25.0	*ns*	5,499.6	*ns*	5.3	*ns*	409.9	*ns*	19.8	*ns*	22.2	4.58**
A × E[a]	2	—	—	—	—	2,821.0	*ns*	.2	*ns*	—	—	11.2	*ns*	—	—
B × Conditions	51	11.0		14.3		2,549.4		2.8		169.7		6.25		4.8	
Total	119														

[a] Stimuli sets are not sources of variance in color naming, repeated objects, and the rod-and-frame test, hence, for these tests $df = 54$ in B × Conditions.

* $p < .06$.
** $p < .05$.
*** $p < .01$.
**** $p < .001$.

TABLE 3
MEANS OF DEPENDENT VARIABLES

Variable	Overlearned repetitive tasks		Perceptual-restructuring tasks		Choice-of-response tasks		EEG
	Color naming (seconds)	Object naming (seconds)	Embedded Figures (seconds)	WAIS Block Designs (raw scores)	Rod and frame (degrees)	Rorschach (no. responses)	Photic driving (no. responses)
Age (A)							
Old	52.02	54.57	137.5	21.42	51.35	4.02	9.43
Middle	49.10	51.77	129.5	21.12	45.20	4.08	8.75
Young	59.50	61.05	191.7	19.47	35.93	3.38	12.65
Time (B)							
a.m.	52.93	54.98	165.9	20.45	46.44	3.30	9.67
p.m.	54.15	56.62	139.9	20.90	41.88	4.35	10.89
Order							
1st	54.91	58.00	173.6	20.16	46.53	3.90	10.10
2nd	52.16	53.60	132.2	21.18	41.79	3.75	10.45
A × B							
Old							
a.m.	51.40	54.40	140.2	21.35	55.60	3.40	8.50
p.m.	52.65	54.75	134.8	21.50	47.10	4.65	10.37
Middle							
a.m.	49.30	50.85	153.2	20.70	47.20	3.65	8.75
p.m.	48.90	52.70	105.7	21.55	43.20	4.50	8.75
Young							
a.m.	58.10	59.70	204.2	19.30	36.53	2.85	11.75
p.m.	60.90	62.40	179.2	19.65	35.35	3.90	13.56
Total *M*	53.54	55.80	152.9	20.68	44.16	3.83	

1953), and one might expect, at a commonsense level, that fatigue would also impair perceptual-restructuring tasks. The concept of a central shift from adrenergic to cholinergic dominance, with this shift having different implications for automatization versus perceptual-restructuring tasks, seems to be an improvement over the fatigue concept in this respect.

However, the basis of the hypothesized shift from central adrenergic toward cholinergic dominance then needs to be explained. A hypothesis for this shift in males has been offered (Broverman et al., 1968; Klaiber, Broverman, Vogel, Abraham, & Cone, 1971). Briefly, this hypothesis suggests that the shift is due to known diurnal rhythms in plasma testosterone levels (Resko & Eik-Nes, 1966). Plasma testosterone concentrations are highest in the morning and slowly subside from that peak throughout the day.

Anthropometric indices of testosterone stimulation (for example, pubic hair development, chest and biceps circumferences) have been reported to be positively related to performances of automatized tasks and negatively related to performances of perceptual-restructuring tasks in males (Broverman, Broverman, Vogel, & Palmer, 1964; Klaiber et al., 1967; Petersen, 1973). Infusions of testosterone (Klaiber et al., 1971; Vogel, Broverman, Klaiber, Abraham, & Cone, 1971) and injections of testosterone (Stenn, Klaiber, Vogel, & Broverman, 1972) have positively affected performances of automatized tasks and EEG indices thought to reflect central adrenergic functioning. Hence, the known morning-to-afternoon decline in testosterone could account for the hypothesized shift in adrenergic versus cholinergic dominance and for the observed changes in performances and in the EEG.

The method by which testosterone may achieve this effect upon brain function has also been investigated. Briefly, both testosterone and estrogen are thought to influence adrenergic functioning in the brain through their influence upon the enzyme monoamine

oxidase (Broverman et al., 1968). Monoamine oxidase activity plays an important role in the intraneural regulation of monoamines believed to be neurotransmitters in adrenergic nerves (Kopin, 1964). Excessive monoamine oxidase activity could result in diminished neurotransmitter levels and impaired central adrenergic functioning. Depressed individuals have been reported to have grossly elevated levels of plasma (Klaiber, Broverman, Vogel, Kobayashi, & Moriarty, 1972) and platelet (Nies, Robinson, Ravaris, & Davis, 1971) monoamine oxidase activity. Monoamine oxidase inhibitors are often used in the treatment of depression (Crane, 1970). Testosterone appears to inhibit monoamine oxidase activity; that is, hypogonadal boys had elevated monoamine oxidase activity compared to normal boys and testosterone therapy in the hypogonadal boys resulted in a shift towards normal levels of monoamine oxidase activity (Klaiber, Broverman, Vogel, & Kobayashi, 1974). Further, plasma monoamine oxidase activity has been found negatively correlated with performances of automatized tasks and positively correlated with performances of perceptual-restructuring tasks (Klaiber et al., 1967).

Hence, the differential shifts in performances of automatized and perceptual-restructuring tasks from a.m. to p.m. could be due to the decline in testosterone levels known to occur during this time. The drop in testosterone, in turn, would permit increased brain monoamine oxidase activity, a decline of available neurotransmitters in adrenergic nerves, impaired adrenergic functioning, and a shift towards cholinergic dominance.

The lack of significant findings involving age is interesting. Apparently, the differences in maturity, which are known to affect level of testosterone production, do not affect diurnal shifts in the measured performances.

It is interesting to speculate on the pedagogical implications of these results, at least for males. For instance, teaching procedures that seem to involve automatized behaviors, for example, drills oriented toward skill acquisition via sustained repetitive practice, might best be conducted in the morning. Similarly, such content areas as speed and accuracy of reading, spelling, and simple additions and subtractions, which have been found to be related to automatization ability (Broverman, 1964; Mathewson, 1967), might be most effectively taught in the morning.

On the other hand, subjects that seem to require cognitive-restructuring behavior might best be taught in the afternoon. Proficiency in abstract thinking, mathematics, and in technical science courses, which have been related to perceptual-restructuring ability (Smith, 1964), might be such subjects. However, more work is needed to determine precisely which classroom subjects and behaviors correlate with, or involve, automatization and perceptual-restructuring abilities.

Unfortunately, the present study cannot be generalized to females. Again, more work is needed.

REFERENCES

Bartley, S. H. Fatigue and inadequacy. *Physiological Reviews*, 1957, **37**, 301–324.

Broverman, D. M. Generality and behavioral correlates of cognitive styles. *Journal of Consulting Psychology*, 1964, **28**, 487–500.

Broverman, D. M., Broverman, I. K., Vogel, W., & Palmer, R. D. The automatization cognitive style and physical development. *Child Development*, 1964, **35**, 1343–1359.

Broverman, D. M., & Klaiber, E. L. Negative relationships between abilities. *Psychometrika*, 1969, **34**, 5–20.

Broverman, D. M., Klaiber, E. L., Kobayashi, Y., & Vogel, W. Roles of activation and inhibition in sex differences in cognitive abilities. *Psychological Review*, 1968, **75**, 23–50.

Crane, G. E. Use of monoamine oxidase inhibiting antidepressants. In W. G. Clark & J. del Giudice (Eds.), *Principles of psychopharmacology*. New York: Academic Press, 1970.

Floyd, W. F., & Welford, A. T. (Eds.) *Symposium on fatigue*. London: Lewis & Co., 1953.

Hollingworth, H. L. Variations in efficiency during the working day. *Psychological Review*, 1914, **21**, 473–491.

Klaiber, E. L., Broverman, D. M., & Kobayashi, Y. The automatization cognitive style, androgens, and monoamine oxidase. *Psychopharmacologia* (Berl.), 1967, **11**, 320–336.

Klaiber, E. L., Broverman, D. M., Vogel, W., Abraham, G. E., & Cone, F. L. Effects of infused testosterone on mental performances and serum

LH. *The Journal of Clinical Endocrinology and Metabolism*, 1971, **33**, 341–349.

KLAIBER, E. L., BROVERMAN, D. M., VOGEL, W., & KOBAYASHI, Y. Effects of depo-testosterone administration upon plasma MAO activity in male patients. Manuscript submitted for publication, 1974.

KLAIBER, E. L., BROVERMAN, D. M., VOGEL, W., KOBAYASHI, Y., & MORIARTY, D. Effects of estrogen therapy on plasma MAO activity and EEG driving responses of depressed women. *American Journal of Psychiatry*, 1972, **128**, 1492–1498.

KOPIN, I. J. Storage and metabolism of catecholamines: The role of monoamine oxidase. *Pharmacological Review*, 1964, **16**, 179–191.

MATHEWSON, S. Oral reading problems in children: A cognitive style approach. Unpublished master's thesis, McGill University, 1967.

MUSCIO, B. Fluctuations in mental efficiency. *British Journal of Psychology*, 1920, **10**, 327–344.

NIES, A., ROBINSON, A. S., RAVARIS, C. L., & DAVIS, J. M. Amines and monoamine oxidase in relation to aging and depression in man. *Psychosomatic Medicine*, 1971, **33**, 470. (Abstract)

PETERSON, A. C. The relationship of androgenicity in males and females to spatial ability and fluent production. Unpublished doctoral dissertation, University of Chicago, 1973.

PHILLIPS, L., KADEN, S., & WALDMAN, M. Rorschach indices of developmental level. *Journal of Genetic Psychology*, 1959, **94**, 267–285.

PHILLIPS, L., & SMITH, J. G. *Rorschach interpretation: Advanced technique*. New York: Grune & Stratton, 1953.

PODELL, J. E., & PHILLIPS, L. A developmental analysis of cognition as observed in dimensions of Rorschach and objective test performances. *Journal of Personality*, 1959, **27**, 439–463.

RESKO, J. A., & EIK-NES, K. B. Diurnal testosterone levels in peripheral plasma of human male subjects. *Journal of Clinical Endocrinology and Metabolism*, 1966, **26**, 573.

SMITH, I. M. *Spatial ability*. San Diego, Calif.: Robert R. Knapp, 1964.

STENN, P. G., KLAIBER, E. L., VOGEL, W., & BROVERMAN, D. M. Testosterone effects upon photic stimulation of the electroencephalogram (EEG) and mental performance of humans. *Perceptual and Motor Skills*, 1972, **34**, 371–378.

STROOP, J. R. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 1935, **18**, 643–672.

VOGEL, W., BROVERMAN, D. M., KLAIBER, E. L., ABRAHAM, G., & CONE, F. L. Effects of testosterone infusions upon EEGs of normal male adults. *Electroencephalography and Clinical Neurophysiology*, 1971, **31**, 400–403.

VOGEL, W., BROVERMAN, D. M., KLAIBER, E. L., & KUN, K. J. EEG responses to photic stimulation as a function of cognitive style. *Electroencephalography and Clinical Neurophysiology*, 1969, **27**, 186–190.

WECHSLER, D. *Wechsler Adult Intelligence Scale*. New York: Psychological Corporation, 1955.

WITKIN, H. A. Individual differences in ease of perception of Embedded Figures. *Journal of Personality*, 1950, **19**, 1–15.

WITKIN, H. A., LEWIS, H. B., HERTZMAN, M., MACHOVER, K., MEISSNER, P. B., & WAPNER, S. *Personality through perception*. New York: Harper, 1954.

(Received March 16, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 247-254

EFFECTS OF FEEDBACK, LEARNER CONTROL, AND COGNITIVE ABILITIES ON STATE ANXIETY AND PERFORMANCE IN A COMPUTER-ASSISTED INSTRUCTION TASK[1]

JOE B. HANSEN[2]

University of Texas at Austin

Ninety-eight undergraduate education majors received a battery of ability tests measuring general reasoning, associative memory, and trait anxiety (A-Trait) and were randomly assigned to three groups—no feedback, feedback, and learner-controlled feedback—for a computer-assisted instruction course on Xenograde systems. State anxiety (A-State) measures were taken (*a*) prior to the course, (*b*) following the administration of stress instructions, (*c*) at the mid-point, and (*d*) at the end. Learner-controlled feedback subjects decreased more in A-State than did feedback subjects. High A-State subjects made more errors under feedback than under no feedback. Feedback facilitated performance for high-reasoning subjects but impaired performance for low-reasoning subjects.

Research on ability-by-treatment interactions has attempted, through the manipulation of task variables, to produce aptitude-by-treatment interactions by altering the relationship between the task and one or more specific abilities known to be important to task performance. Such studies have sought to establish principles that would lead to the development of instructional design models more sensitive to individual learner differences. A separate, equally important domain of research has dealt with motivational factors in learning as an approach to the general problem of individualization of instruction. Cattell (1966) has suggested that anxiety is a function of unresolved doubt about an expected outcome. If so, then providing feedback could reduce subjects' doubt about performance on the learning task and, consequently, reduce anxiety. Drive theory (Spence & Spence, 1966) predicts that high-anxiety subjects will perform better than low-anxiety subjects on simple tasks in which a single habit tendency is involved, while low-anxiety subjects should be superior on complex tasks in which competing habit tendencies are involved.

Spielberger (1966) distinguishes Trait-Anxiety (A-Trait), a relatively permanent personality variable, from State-Anxiety (A-State), a transitory condition resulting from the amount of threat perceived by an individual in a particular situation. The State-Trait Anxiety Inventory was developed by Spielberger, Gorsuch, and Lushene (1970) as a means of measuring these two types of anxiety separately.

The major objectives of this study were (*a*) to determine whether information feedback provided during a difficult task will reduce A-State and (*b*) to determine whether learner control of feedback will lead to further reductions in A-State. Another objective was to attempt to bridge cognitive and affective domains by examining the relationships between the task variables of feedback and learner control

[1] This study was supported by National Science Foundation Contract GJ 509X.

[2] Requests for reprints should be sent to Joe B. Hansen, who is now with the Evaluation Department, Portland Public Schools, 631 Northeast Clackamas Street, Portland, Oregon 97208.

with anxiety and cognitive abilities, specifically reasoning ability and associative memory.

The following specific hypotheses were investigated in this study:

1. Subjects who receive informative feedback and subjects who have learner control over feedback show greater reductions in A-State during a computer-assisted instruction task than subjects who receive no feedback.

2. High A-State subjects produce fewer errors under the feedback condition than under the no-feedback condition, while low A-State subjects produce fewer errors under the no-feedback than feedback condition.

3. Subjects in the no-feedback group show a higher demand for both reasoning and associative memory abilities than subjects in either the feedback or learner control of feedback groups; that is, the amount of change in error score per unit of increase in each of the ability scores would be greater in absolute terms for the no-feedback subjects than for subjects in either of the other groups.

Method

Subjects

The subjects were 98 undergraduate female education majors at the University of Texas at Austin who were randomly assigned to three groups: no feedback ($n = 38$), feedback ($n = 31$), and learner-controlled feedback ($n = 29$).

Apparatus

The task, a computer-assisted instruction course on the artificial science, Xenograde (Merrill, 1964), was a revision of an earlier version described in detail by Merrill (1970) and Bunderson and Hansen (1972). It was presented by means of an IBM 1500 computer system in the computer-assisted instruction laboratory of the University of Texas at Austin. The system has eight terminals of the cathode ray tube type (IBM 1510). Each terminal is accompanied by an image projector (IBM 1512) for the computer-controlled presentation of 16-millimeter transparencies. The terminals, each housed in an individual wooden carrel constructed to provide isolation and work space for each student, are all located in the same room of the computer-assisted instruction laboratory.

Procedure

Approximately two weeks prior to taking the computer-assisted instruction course, each subject received a battery of five tests which were administered to small groups at the convenience of the subjects. The test battery included the A-Trait scale of the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970); the Ship Destinations Test, the Object–Number Test, and the First and Last Names Test (French, Ekstrom, & Price, 1963); and the Bi-Column Number Series Test (Merrill, 1970). The Ship Destinations Test is a measure of general reasoning ability, while the Bi-Column Number Series Test is designed to measure the subject's facility with the information-processing requirements of the task. The Object–Number and the First and Last Names Tests are measures of associative memory.

As each subject reported to the terminal room, she was assigned to a terminal and was immediately given a 20-item paper-and-pencil version of the A-State scale of the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970). Following the A-State measure, subjects received appropriate instructions on terminal operation, followed by ego-involving stress instructions given on-line. The stress instructions implied that the task was an indicator of intelligence and that each subject would be compared with other college students. Following the stress instructions, a 5-item version of the A-State scale of the State-Trait Anxiety Inventory was presented on-line, followed immediately by the first example of the Xenograde course. Figure 1 is a chart showing the task structure.

The course contained a series of eight sets of three examples and three test items illustrating eight consecutive hierarchical rules comprising the task. Following each example, three test questions designed to test the subject's knowledge of the exemplified rule were presented. In the no-feedback group, subjects received no feedback following their test-item responses. In the feedback group, they received the words "true" or "false" as feedback following each test item, plus a statement of the rule following the ninth test item for each rule. In the learner-controlled feedback condition, subjects were required to type "y" to receive the true–false feedback and "n" if no feedback was desired following a test item. The learner-controlled feedback group also received the option of viewing the rule on the image projector following the third test item for any example. Presentation of the rule terminated the presentation of examples for that rule and resulted in immediate presentation of the first example of the next rule.

All subjects received a second five-item A-State scale on-line, following the last test item for the fourth rule, and a third five-item A-State scale following the last test item for the eighth rule. In this manner, A-State changes for each subject were tracked over the task period.

Results and Discussion

A varimax factor analysis conducted on the ability scores produced two factors—

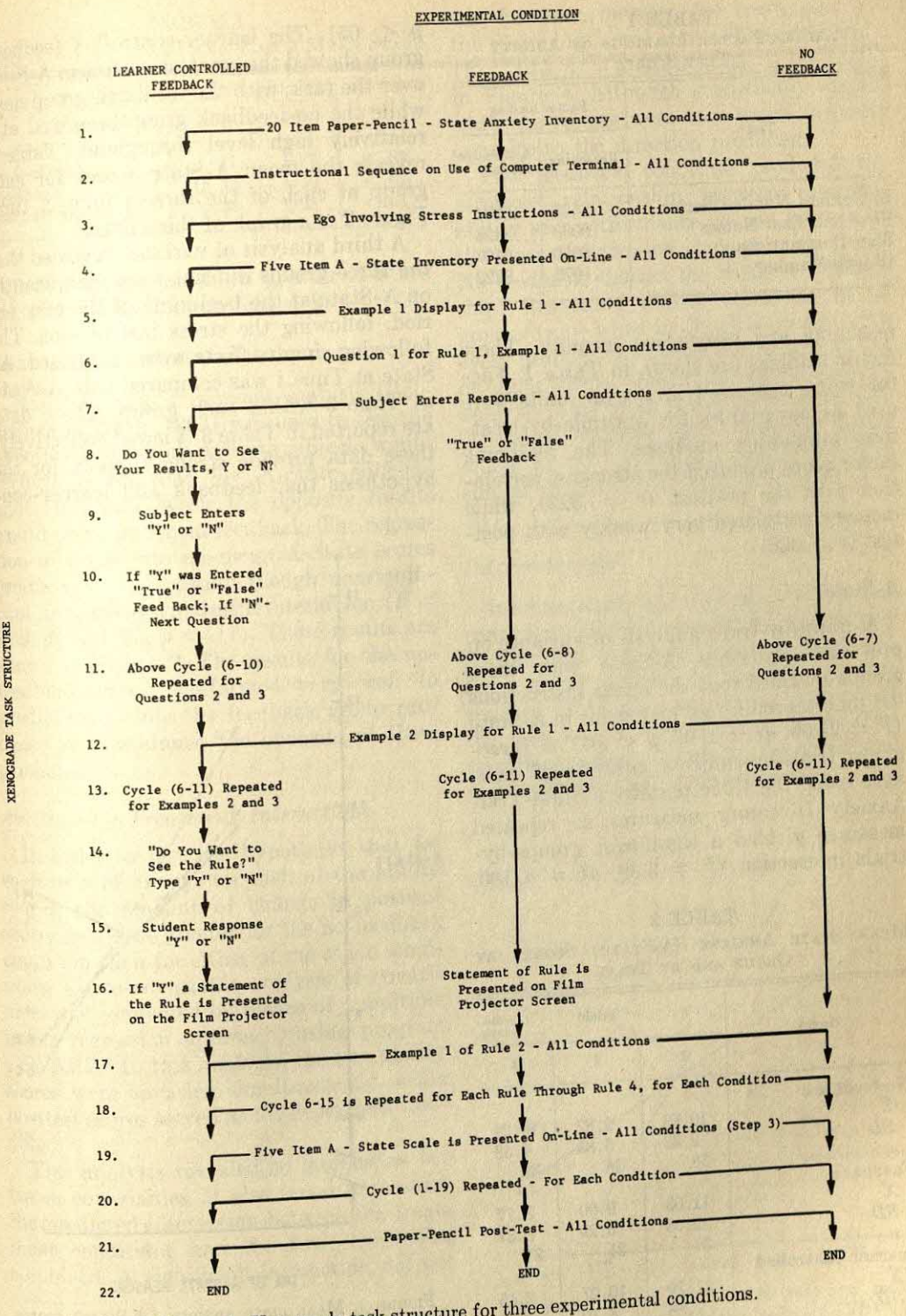


FIGURE 1. Xenograde task structure for three experimental conditions.

TABLE 1
VARIMAX FACTOR LOADINGS OF ABILITY TEST SCORES

Test	Factor loading	
	Reasoning	Associative memory
Bi-Column Number Series	.8525	.0901
First and Last Names	−.0089	.8428
Ship Destinations	.8659	.0107
Object–Number	.1076	.8222

reasoning and associative memory. These factor loadings are shown in Table 1. Factor scores expressed in z score units were used as covariables for aptitude-by-treatment-interaction analyses. The reasoning factor score produced the strongest correlation with the posttest ($r = .322$), while memory correlated very weakly with posttest ($r = .026$).

A-State

A group-by-trials analysis of variance on pre- and poststress A-State with three groups revealed that the stress instructions did indeed produce an increment in A-State ($F = 27.99$, $df = 1/97$, $p < .01$). A three-group trials-by-subjects analysis of variance using the three poststress State-Trait Anxiety Inventory measures as repeated measures yielded a significant groups-by-trials interaction ($F = 3.39$, $df = 4/190$, $p < .05$). The learner-controlled feedback group showed the greatest decline in A-State over the task, with the feedback group next, while the no-feedback group remained at a relatively high level throughout. Table 2 reveals the mean A-State scores for each group at each of the three points in time. Figure 2 is a graph of these data.

A third analysis of variance revealed that the three groups did not vary significantly on A-State at the beginning of the task period, following the stress instructions. The following simple effects were also tested: A-State at Time 1 was compared with A-State at Time 3 within each group. These data are reported in Table 3. Viewed collectively, these data provide strong support for the hypothesis that feedback and learner-con-

TABLE 2
MEAN STATE ANXIETY (A-STATE) SCORES BY GROUP AND BY TRIAL

Group	Trials		
	1	2	3
No feedback			
$\bar{X}$	10.89	9.87	10.24
SD	2.82	3.85	4.53
n	38	38	38
Feedback			
$\bar{X}$	11.03	9.90	8.77
SD	2.68	3.29	3.37
n	31	31	31
Learner controlled			
$\bar{X}$	11.79	10.21	8.45
SD	2.88	3.70	3.05
n	29	29	29

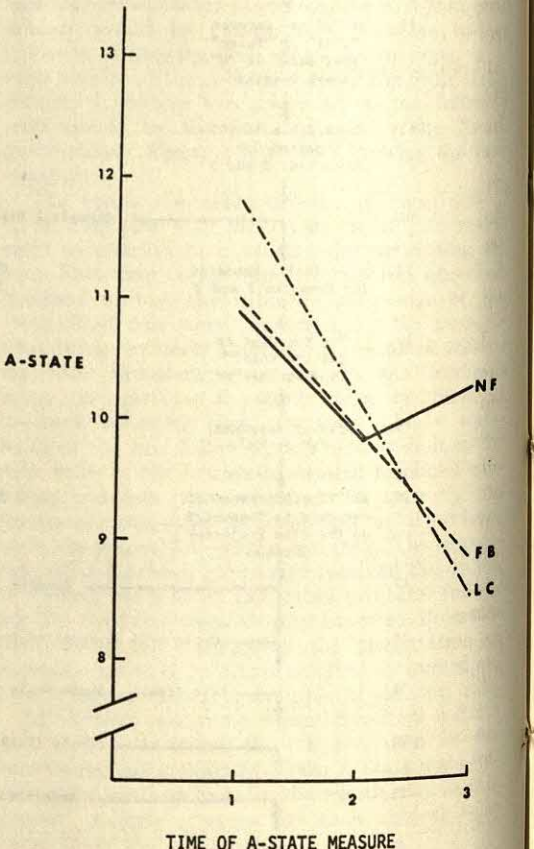


FIGURE 2. Mean state anxiety (A-State) scores by group. (Abbreviations: NF = no feedback, FB = feedback, and LC = learner-controlled feedback.)

TABLE 3
COMPARISON OF TIME 1 AND TIME 3 MEAN STATE ANXIETY (A-STATE) SCORES BY GROUP

Group	F	df	p
No feedback	.77	1/37	$<.39$
Feedback	18.38	1/30	$<.01$
Learner controlled	36.55	1/28	$<.01$

trolled feedback result in greater A-State reductions than does no feedback.

Anxiety-by-Treatment Interactions

Multiple linear regression analysis (Bottenberg & Ward, 1963) was used to test the hypothesis that high A-State subjects would perform better than low A-State subjects under feedback, while the opposite results would occur under no feedback. The regression of error rate on mean A-State scores produced an interesting though nonsignificant anxiety-by-treatment interaction ($F = 2.69$, $df = 1/65$, $p < .11$). These results are shown in Figure 3. The results for the no-feedback group conformed very well to predictions, while the feedback group produced results almost the opposite of those predicted.

Aptitude-by-Treatment Interactions

In order to test the hypothesis that for each unit of change on each of the ability scores, the amount of change in posttest scores would be greater for the no-feedback condition than for either of the other conditions, a two-covariable analysis of covariance was conducted by means of a multiple linear regression approach, using program COVAR2.[3] In this analysis, the two ability scores were covaried simultaneously, while posttest scores served as the dependent variable.

This analysis revealed no interaction between covariables. It also failed to produce the predicted interaction between the treatment conditions and the two covariables combined. A significant interaction did occur, however, between the treatments and the reasoning factor scores. This interaction ($F = 3.15$, $df = 2/91$, $p < .05$) is illustrated in Figure 4. Although statistically significant, the results were once again virtually opposite to the direction predicted.

The feedback condition produced the strongest relationship between reasoning ability and performance, as can be seen by the steeper regression slope for this group. No feedback showed the weakest relationship. Reasoning ability seems to be of greater value when feedback is present after every problem than when feedback is absent or under learner control. A single covariable analysis using mean errors per problem as a dependent variable and reasoning factor scores as the covariables produced highly similar results ($F = 7.27$, $df = 2/92$, $p < .01$).

Interpretations

It seems clear that feedback and learner-controlled feedback resulted in reductions in A-State not obtained under no feedback. Some of the decrease in A-State probably resulted from the subjects' adaptation to the task and to the computer-assisted instruction medium. However, the crossing of the feedback and learner-controlled feed-

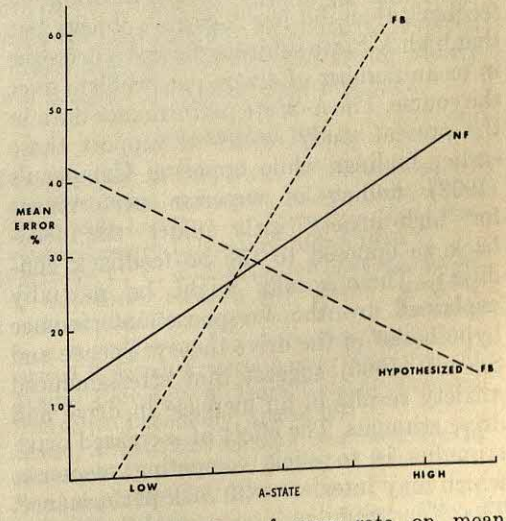


FIGURE 3. Regression of error rate on mean state anxiety (A-State) for the feedback (FB) and no feedback (NF) groups.

[3] This program was written by E. E. Jennings, the University of Texas at Austin (Ward & Jennings, 1973).

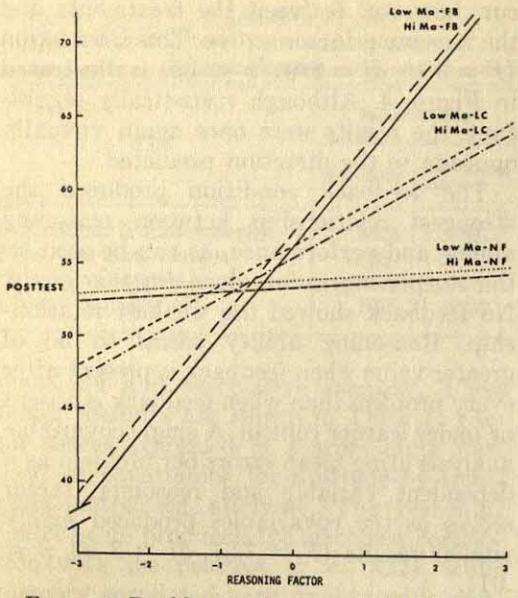


FIGURE 4. Double covariance analysis with posttest as dependent measure and associative memory (Ma) and reasoning factor scores as covariables. (Abbreviations: FB = feedback, LC = learner-controlled feedback, and NF = no feedback.)

back group means over the no-feedback group as shown in Figure 2 suggests that the treatments were effective in reducing A-State.

Spielberger, O'Neil, and Hansen (1970) found that high A-State subjects performed more poorly under conditions of informative feedback than did low A-State subjects but that high A-State subjects showed a decrease in mean number of errors per problem over the course. The A-State performance data in the present study seems to support those earlier findings while opposing Campeau's (1968) finding of superior performance for high-anxiety girls under the feedback as opposed to the no-feedback condition. These results might be partially explained by the "response interference hypothesis" of the drive theory. Spence and Spence (1966) suggest that stress-induced anxiety results in an increase in drive and drive stimulus. The effect of increased drive stimulus is to elicit competing responses which may interfere with task performance. The Xenograde task can be characterized as a hypothesis formation–testing task. If the subjects' response pattern can be defined in terms of a hypothesis formation–evaluation–rejection cycle, then individual hypotheses about what constitutes an appropriate rule can be thought of as individual, covert, mediating responses to example displays. This then leads to overt attempts to solve the ensuing problems, with the attempted solution providing a basis for rejection or acceptance of the hypothesis in question. In such a situation, then, high A-State could result in the generation of a greater number of competing erroneous hypotheses. In order to test these hypotheses, the subject must utilize information available in the problem displays or present in the form of feedback. If we can assume a limit on the amount of information a subject can process—that is, channel capacity, as suggested by Miller (1967)—then it seems reasonable to speculate that increased information input in the form of feedback could contribute to an increase in the proportion of incorrect hypotheses, thereby producing a loss of efficiency in the hypothesis formation–evaluation process. This could account for the greater number of errors per problem under the feedback condition for high A-State subjects.

To summarize, it seems plausible that increased anxiety produces a greater number of competitive responses in the form of erroneous hypotheses. Given an upper limit on information-processing capacity, the subject must now evaluate a higher proportion of erroneous hypotheses. Feedback information further adds to the information-processing burden, resulting in reduced efficiency and in an increase in the error–problem ratio. Although speculative, such reasoning implies the need for further studies in which the effects of high A-State on information-processing variables can be more carefully examined. An approach to such studies is suggested below.

CONCLUSIONS

Several tentative conclusions can be drawn from the results of this study regarding the effects of feedback on A-State and the relationships among A-State, ability, and performance. With respect to the effects of information feedback on A-State, it ap-

pears that real reductions in A-State can be obtained through increased use of feedback. Whether or not this results in higher levels of performance or improved learning depends on other factors, particularly on ability factors that are known to be important to the task. Feedback seems to help persons with high reasoning ability while hindering the performance of those with low reasoning ability; this suggests a positive relationship between reasoning and information-processing capacities.

While feedback generally seems to reduce A-State, high A-State appears to interfere with the learner's capacity to utilize the feedback information effectively in performing the task requirements. Learner control, although defined here in a limited manner, also seems to offer definite advantages both in terms of anxiety reduction and performance. While the learner-controlled feedback was equally effective with the feedback condition in reducing anxiety, it resulted in a substantial reduction in the amount of work required to complete the task.

Suggested Further Research

The suggested relationships among the variables of A-State, information processing, and learning indicate a need for further research. Costello and Dunham (1971) have described a methodology, in the form of the "approach model," which offers promise for the investigation of the relationships between two classes of variables, those relating to task performance and those relating to cognitive processes. The procedure embodied by the approach model typically involves the administration of tests of a mental ability on which there is some general consensus of acceptance, such as induction or associative memory. It also involves the administration of a learning problem—usually a concept-learning task—selected for its suspected ability requirements. The ability tests are then submitted to a "rational information-processing analysis" and further tests are developed. These new tests are expected to be tests of the specific information-processing variables that are inherent in the ability tests. An example appropriate to the reasoning ability factor might be hypothesis generation, hypothesis evaluation, or both of these. An analogous set of tests is developed from a rational information-processing analysis of the task requirements. A factor analysis of the two sets of derived test scores will reveal, through common factor loadings, factors that are inherent to both the task and the ability in question.

The applicability of the approach model is limited to investigation of cognitive processes and therefore would not be of value in investigating the relationship between cognitive and affective processes. It should, however, provide a sound methodology for investigation into the relationship between reasoning and information-processing variables. By introducing varied feedback information content into the task as an independent variable, one might expect differential process requirement structures under different feedback conditions, with the relationship between process measures and performance measures also varying according to feedback condition. Higher feedback may reduce the magnitude of the relationship between performance and Process A while increasing the magnitude of the relationship between performance and Process B. The inclusion of A-State measures at various points in the learning task might provide a means of determining more precisely the nature of the relationship between A-State and information-processing ability as a function of differential feedback.

REFERENCES

Bottenberg, R. A., & Ward, J. H., Jr. *Applied multiple linear regression.* (Tech. Rep. PRL-TDR-63-6) Lackland Air Force Base, Tex.: Personnel Research Laboratory, March 1963.

Bunderson, C. V., & Hansen, J. B. *The interaction of associative memory and general reasoning with availability and complexity of examples in a computer-assisted instruction task.* (Advanced Research Projects Agency Tech. Rep. 6) Austin: University of Texas, Computer-Assisted Instruction Laboratory, August 1971.

Campeau, P. L. Test anxiety and feedback in programmed instruction. *Journal of Educational Psychology,* 1968, **59,** 159–163.

Cattell, R. B. Anxiety and motivation: Theory and crucial experiments. In C. D. Spielberger (Ed.), *Anxiety and behavior.* New York: Academic Press, 1966.

COSTELLO, R. J., & DUNHAM, J. L. Inductive reasoning processes in concept learning. Paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. *Manual for Kit of Reference Tests for Cognitive Factors.* Princeton, N.J.: Educational Testing Service, 1963.

MERRILL, M. D. *Transfer effects within a hierarchical task as a function of review and correction on successive parts.* (Tech. Rep. 5, Navy/Office of Naval Research Contract 3985(04)) Urbana, Ill.: Training Research Laboratory, 1964.

MERRILL, P. F. *Interaction of cognitive abilities with availability of behavioral objectives in a hierarchical task by computer-assisted instruction.* (Tech. Rep. 5, Office of Naval Research Contract N00014-67-A-0126-0006) Austin: University of Texas, May 1970.

MILLER, G. A. *Psychology of Communication.* Baltimore, Md., Penguin Books, 1967.

SPENCE, J. T., & SPENCE, K. W. The motivational components of manifest anxiety: Drive and stimuli. In C. D. Spielberger (Ed.), *Anxiety and behavior.* New York: Academic Press, 1966.

SPIELBERGER, C. D. Theory and research on anxiety. In C. D. Spielberger (Ed.), *Anixety and behavior.* New York: Academic Press, 1966.

SPIELBERGER, C. D., GORSUCH, R. L., & LUSHENE, R. E. *Manual for the State-Trait Anxiety Inventory.* Palo Alto, Calif.: Consulting Psychologist Press, 1970.

SPIELBERGER, C. D., O'NEIL, H. F., JR., & HANSEN, D. N. *Anxiety, drive theory and computer-assisted learning.* (Tech. Rep. 14) Tallahassee: Florida State University, Computer-Assisted Instruction Center, 1970.

WARD, J. H., & JENNINGS, E. E. *Introduction to linear models.* Englewood Cliffs, N.J.: Prentice-Hall, 1973.

(Received March 21, 1973)

Manuscripts Accepted for Publication in the *Journal of Educational Psychology*

Effects of Nonparticipant Observation on Teacher and Pupil Classroom Behavior. Michael Mercatoris and W. Edward Craighead (Department of Psychology, The Pennsylvania State University, 512 Moore Building, University Park, Pennsylvania 16802).

Sensitivity and Validity of Learning Potential Measurement in Three Levels of Ability. Elisha Y. Bahad and Milton Budoff (Research Institute for Educational Problems, 12 Maple Avenue, Cambridge, Massachusetts 02139).

Instructional Objectives as Directions to Learners: Effect of Passage Length and Amount of Objective–Relevant Content. Robert Kaplan (Bell Telephone Laboratories, CPI L101, P.O. Box 2020, New Brunswick, New Jersey 08903) and Ernst Z. Rothkopf.

Cognitive Style and Reasoning about Speed. Linnea C. Ehri (Department of Education, University of California, Davis, California 95616) and Irene M. Muzio.

Reasoning and Praise: Their Effects on Academic Behavior. Suzanne J. Taffel, K. Daniel O'Leary (Department of Psychology, State University of New York, Stony Brook, New York 11790), and Sandra Armel.

Male and Female Teacher Attitudes as a Function of Students' Ascribed Motivation and Performance Levels. Larry J. Brandt (Department of Psychology, University of Houston, Houston, Texas 77004) and Mary Ellen Hayden.

Teacher Behavior Across Ability Groups: A Consideration of the Mediation of Pygmalion Effects. Judith Landon Alpert (175 West Twelfth Street, New York, New York 10011).

Algorithmic Approach to Curriculum Construction: A Field Test in Mathematics. Walter Ehrenpreis and Joseph M. Scandura (Graduate School of Education, University of Pennsylvania, Philadelphia, Pennsylvania 19104).

Internalization of Filmic Schematic Operations in Interaction with Learners' Aptitudes. Gavriel Salomon (School of Education, The Hebrew University, Jerusalem, Israel).

Changing Teacher and Student Behavior: An Empirical Investigation. Thomas L. Good (Center for Research in Social Behavior, University of Missouri, Columbia, Missouri 65201) and Jere E. Brophy.

Some Psychometric Questions in the Evaluation of Professors. Richard M. Jaeger (College of Education, University of South Florida, Tampa, Florida 33620) and Tom D. Freijo.

Origins of Teachers' Attitudes toward Young Children. Sherry L. Willis (College of Human Development, Pennsylvania State University, University Park, Pennsylvania 16802) and Jere Brophy.

Journal of Educational Psychology
1974, Vol. 66, No. 2, 255–261

A TEST OF THE THEORY OF FLUID AND CRYSTALLIZED INTELLIGENCE IN MIDDLE- AND LOW-SOCIOECONOMIC-STATUS CHILDREN:

A CROSS-LAGGED PANEL ANALYSIS

FRANK L. SCHMIDT AND WILLIAM D. CRANO[1]

Michigan State University

The prediction from the theory of fluid and crystallized intelligence that fluid ability (g_f) operates as a cause of crystallized ability (g_c) was tested by means of the cross-lagged correlation panel technique in samples of middle- (n = 3,944) and low-socioeconomic-status (n = 1,501) elementary schoolchildren. Support for the postulated causal relationship was found in the middle-socioeconomic-status group, but results for the low-socioeconomic-status group did not support the prediction. Possible explanations for this disparity were presented and discussed.

The theory of fluid and crystallized intelligence, developed and elaborated primarily by Cattell (1971) and Horn (1968), holds that there is not one but two general intelligences: fluid ability (g_f) is conceptualized as abstract, essentially nonverbal, relatively culture-free mental efficiency, while crystallized mental ability (g_c) is viewed as consisting primarily of acquired skills and knowledges and is thus strongly dependent on cultural exposures. The g_f is involved heavily in performance on such nonverbal tests as figure classification, figural analogies, and number and letter series tests and matrices. On the other hand, tests of vocabulary, general information, abstruse word analogies, and the mechanics of language are considered to be almost pure measures of g_c. (Such tests as arithmetic reasoning, inductive verbal reasoning, and syllogistic reasoning load about equally on the g_f and g_c factors.) The g_c is viewed as the "extent to which one has appropriated the collective intelligence of his culture for his own use [Horn & Cattell, 1967, p. 111]," while g_f is said to reflect, to a much greater extent, basic neurophysiological efficiency. An impressive amount of empirical evidence has accumulated to support this distinction (Cattell, 1963, 1971; Horn & Cattell, 1966a, 1966b, 1967).

This theory postulates that in the young child, g_f stands in a causal relation to g_c. In addition to such nonintellective causes as educational opportunities and individual motivation, acquisition of g_c depends on the level of g_f (Cattell, 1971, pp. 98–100, 117–118; Horn, 1968).[2] However, the reverse causal relation is not a part of the theory; level of g_f is considered to be unaffected by previous g_c acquisitions. This pattern of causality is not predicted by competing theories of intelligence. Although the hierarchical group-factor theories advanced by Vernon (1961), Burt (1949, 1955) and others do contain a pair of factors (verbal-educational, and practical–mechanical–spatial) that are superficially similar to g_c and g_f, they postulate no causal relations between the two.[3] Similarly, neither Thurstone's (1938) nor Guilford's (1967) theory contains any such causal propositions. Thus, empiri-

[1] Requests for reprints should be sent to William D. Crano, Department of Psychology, Michigan State University, East Lansing, Michigan 48824.

[2] Thus a moderate correlation (.40 to .50) would be expected, and is found, between the two general ability factors (Cattell, 1971, pp. 98–100).

[3] For a description of the differences, see Horn (1968) and Cattell (1971, pp. 100–102).

cal verification of the differential causal prediction would lend strong support to the Horn-Cattell (1966b) position.

The purpose of the present study is to test this causal prediction separately in samples of middle- and low-socioeconomic-status elementary schoolchildren using the cross-lagged panel correlation technique (Campbell, 1963).

Method

Subjects

Subjects were 5,495 students enrolled in the public schools of Milwaukee who had been administered the Lorge-Thorndike Intelligence Test (1957 version, Level 3) and the Iowa Tests of Basic Skills in Grade 4 and alternate forms of these same tests in Grade 6 during the academic years 1963–1964 and 1965–1966, respectively. Students from schools eligible for comprehensive programs of aid under Title 1 of the Elementary and Secondary Education Act for the 1967–1968 school year were classified as low socioeconomic status (n = 1,501); students from ineligible schools were considered essentially middle socioeconomic status.

The g_f and g_c Measures

The Lorge-Thorndike Intelligence Test (Lorge, Thorndike, & Hagan, 1966) contains both a Verbal and a Nonverbal scale. The Lorge-Thorndike Verbal scale is inappropriate to the purposes of this study because two of its five subtests tap abilities which can be expected to have a large loading on g_f: Arithmetic Reasoning (Subtest 3) and Analogies Using Common Words (Subtest 5; Cattell, 1971, p. 97 ff.; Horn, 1968; Horn & Cattell, 1967). The Lorge-Thorndike Verbal scores are best considered composites of g_c and g_f.[4] The Lorge-Thorndike Nonverbal, on the other hand, is apparently a nearly pure measure of g_f. Its three subtests—Figure Analogies, Figure Classification, and Number Series—are entirely pictorial, diagrammatic, or numerical in nature. Each of these kinds of measures has been reported as loading heavily on g_f and negligibly or not at all on g_c (Cattell, 1971, p. 97 ff.; Horn, 1968; Horn & Cattell, 1967). Additional evidence for the construct validity of Lorge-Thorndike Nonverbal as a measure of g_f is provided in the Lorge et al. (1966) manual: In two separate samples in which the Lorge-Thorndike Nonverbal was correlated with the tests in the Differential Aptitude Tests, its highest correlations (.76 and .63) were with the Abstract Reasoning scale, a nonverbal test which fits Cattell's (1971, pp. 98–99) description of the ideal g_f measure almost perfectly.

[4] Scores are not reported separately for subscales within the Verbal and Nonverbal scales.

The Iowa Test of Basic Skills (Lindquist & Hieronymus, 1964) consists of five subscales:

1. Vocabulary;
2. Reading Comprehension;
3. Language, which is made up of the subscales Spelling, Capitalization, Punctuation, and Language Usage;
4. Work-Study Skills, which includes the subscales Map Reading, Graph and Table Reading, and Knowledge and Use of References; and
5. Arithmetic, which consists of the subscales Arithmetic Concepts and Arithmetic Problem Solving.[5]

Vocabulary, Reading Comprehension, Spelling, Capitalization, Punctuation, Language Usage (detection of grammatical errors), and Knowledge and Use of References are clearly measures of g_c and were employed as such in this study. Map Reading and Graph and Table Reading probably load on general visualization ability (g_v), which is considered as a separate factor in the theory of g_f and g_c.[6] On initial consideration, the Arithmetic Concepts subscale, which assesses mastery of the concepts involving the number system, decimals, ratios, etc., appears to be an almost pure measure of g_c, but examination of individual items in this measure revealed that many involve problem solving and general reasoning, making it likely that this scale taps the g_f factor to some extent. Available evidence (Cattell, 1971, p. 97 ff.; Horne, 1968; Horn & Cattell, 1967) indicates that arithmetic reasoning (solution of word problems) loads about equally on g_f and g_c and thus the Arithmetic Problem-Solving subscale was also eliminated as a g_c measure, leaving seven relatively pure measures of g_c.

Analysis

The analytic procedure employed in this study was the cross-lagged panel correlational technique (Campbell, 1963; Campbell & Stanley, 1963; Pelz & Andrews, 1964). When one has available correlational information relating two variables at two or more points in time, this technique can allow the inference of causality. This inference is based on the supposition of all sciences that when a given event (Event 1) consistently precedes the occurrence of another (Event 2), but the opposite does not hold, only two conclusions are possible: (*a*) Event 1 is a cause (possibly only one of many) of Event 2 or (*b*) both Events 1 and 2 are the effects of some more general cause or causes. In experimental designs, control of the presentation of the independent variable is seen to rule out the second alternative, and thus, response differences be-

[5] For a fuller description of each of these subscales see Lindquist and Hieronymus (1964).

[6] Factors regarded by the theory as separate from (and much less important than) g_f and g_c include general visualization (g_v), general speediness (g_s), carefulness (C), and fluency (F).

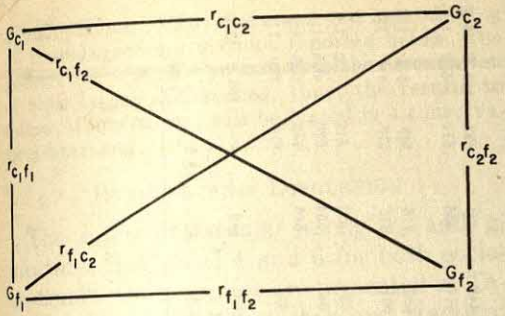


FIGURE 1. Schematic representation of associated relationships in a cross-lagged panel analysis investigating fluid (f) and crystallized (c) intelligence (G).

tween experimental and control subjects are taken as effects of the experimental manipulation or treatment. Correlational studies, of course, do not typically entail the controlled application of an independent manipulation. Both variables are, in effect, dependent measures. Hence, causal inference on the basis of correlational results is unsupportable, given the possibility of the second alternative noted above. Consider, however, the pattern of results obtained in a simple cross-lagged panel analysis (Figure 1). For purposes of later exposition, the two hypothetical variables employed will consist of fluid (f) and crystallized (c) intelligence measured at two time periods. The two synchronous correlations ($r_{c_1f_1}$, $r_{c_2f_2}$) and the lagged autocorrelations ($r_{c_1c_2}$, $r_{f_1f_2}$) provide information concerning the relationship between the two constructs, and the temporal stability of the tests employed in the measurement of these constructs.

More central from the standpoint of causal inference are the correlations crossed and lagged over time (i.e., $r_{c_1f_2}$, $r_{f_1c_2}$). Suppose, for example, that status on g_f at Time 1 was followed consistently by a similar status in g_c at Time 2 but that the opposite did not hold. In that case, $r_{f_1c_2} > r_{c_1f_2}$, and on the basis of the time-precedence notions of causality discussed above, one could argue that g_f was a cause of later g_c. If $r_{c_1f_2} > r_{f_1c_2}$, of course, the opposite causal inference would be supported. In the null case in which $r_{f_1c_2} = r_{c_1f_2}$, either no causal relationship exists or both variables are the effect of some hidden cause (see Crano & Brewer, 1973).

In the present study, the prediction is that with appropriate statistical corrections for reliability and communality changes (see Crano, et al., 1972), the crossed and lagged correlation $r_{f_1c_2}$ (i.e., the correlation between g_f in Grade 4 and g_c in Grade 6) will be significantly larger than $r_{c_1f_2}$ (the correlation between g_c in Grade 4 and g_f in Grade 6) for each of the seven measures of g_c. If the predominant direction of causation is from g_f to g_c rather than vice versa, it would be expected that the correspondence between relative standing on g_f in Grade 4 and g_c in Grade 6 would be greater than the correspondence between g_c in Grade 4 and g_f in Grade 6. Confirmation of this hypothesis would not eliminate the possibility that g_c to some extent influences g_f, but it would indicate that the preponderance of causation is in the direction of g_f causing g_c.[7] Another point to be borne in mind is that, as noted above, g_f is postulated as one of a *number* of causes of g_c. Thus, confirmation of the present hypothesis would provide evidence about only one g_c determinant.

As with most powerful methodologies, the cross-lagged panel correlational technique is not without its technical problems. One of the most difficult of these comes about as a result of the time lag which occurs between the administrations of the tests of interest. Over the course of time, it is conceivable that test characteristics (e.g., common factor structure and internal consistency) might change and, further, that these changes would indicate a significant discontinuity in the cross-lagged correlations even in the absence of a true causal relationship. For example, suppose that the reliability of the g_f measure increased (or its specificity decreased) over time, while the reliability of the g_c measure decreased (or its specificity increased). The result would be an increase in $r_{c_1f_2}$ and a decrease in $r_{f_1c_2}$; these effects might be large enough to indicate a significant disparity between the cross-lagged correlations in which no causal relationship existed; such effects could be of a magnitude large enough to reverse the true direction of causality.

In attempting to offset this problem, one could adopt a factor analytic approach that allows for variation in the unique factor loading of each variable (over time) but assumes that all orthogonal common factor loadings are invariant or that they change by some multiplicative constant. If this assumption is met, then the common factor structure of the tests over time would be unchanged, although there might well be changes in communality and, therefore, uniqueness. Having corrected for the effects of this change constant on the cross-lagged correlations, one could inspect the resulting causal discontinuities without undue concern for the potentially biasing effects of unreliability and factor structure changes.

To test the viability of this assumption, all synchronous Time 1 correlations (e.g., $r_{f_1c_1}$) could be divided by their respective Time 2 correlations ($r_{f_2c_2}$). If the assumption of a multiplicative constant change in common factor structure is valid, then the resulting matrix of ratios will be single factored. Within the bounds of sampling error, such a criterion was met in this research. The correction procedure (the specifics of which are detailed in Crano, Kenny, & Campbell, 1972, pp.

[7] A fully adequate test of the theory of g_f and g_c would have to be capable of eliminating this possibility, since the apparent prediction from the theory is that g_c cannot operate causally with respect to g_f.

TABLE 1

MATRIX OF INTERCORRELATIONS FOR MIDDLE-SOCIOECONOMIC-STATUS GROUP (UPPER RIGHT TRIANGULAR MATRIX) AND LOWER-SOCIOECONOMIC-STATUS GROUP (LOWER LEFT TRIANGLE) WITH MEAN GRADE EQUIVALENT OR IQ SCORE AND STANDARD DEVIATION FOR EACH MEASURE

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	M	σ
1. Vocabulary (4)	—	.731	.579	.565	.457	.682	.581	.465	.744	.691	.588	.555	.530	.637	.620	.471	4.4	.90
2. Reading Comprehension (4)	.623	—	.575	.596	.514	.660	.629	.511	.693	.711	.571	.560	.532	.612	.644	.492	4.2	.97
3. Spelling (4)	.533	.502	—	.631	.540	.559	.546	.410	.488	.485	.735	.562	.535	.512	.574	.398	4.5	1.1
4. Capitalization (4)	.537	.560	.560	—	.628	.576	.591	.476	.498	.507	.566	.619	.546	.522	.609	.469	4.3	.90
5. Punctuation (4)	.450	.488	.451	.611	—	.524	.503	.424	.422	.440	.504	.545	.539	.457	.527	.413	4.3	1.0
6. Usage (4)	.564	.557	.511	.556	.497	—	.555	.451	.649	.617	.570	.569	.560	.692	.589	.451	4.2	1.2
7. References (4)	.455	.484	.453	.468	.416	.435	—	.462	.526	.539	.536	.523	.498	.496	.603	.453	4.4	.88
8. Nonverbal IQ (4)	.450	.424	.424	.477	.418	.418	.423	—	.469	.502	.380	.475	.452	.448	.461	.692	101.5	14.7
9. Vocabulary (6)	.595	.548	.412	.443	.416	.562	.397	.401	—	.809	.582	.577	.547	.676	.656	.511	6.2	1.4
10. Reading Comprehension (6)	.610	.594	.432	.508	.455	.566	.449	.486	.727	—	.575	.601	.574	.656	.716	.549	6.3	1.2
11. Spelling (6)	.518	.485	.680	.498	.425	.516	.436	.404	.527	.535	—	.651	.611	.601	.645	.430	6.4	1.3
12. Capitalization (6)	.488	.483	.481	.581	.516	.518	.401	.450	.527	.574	.616	—	.705	.624	.681	.516	6.9	1.6
13. Punctuation (6)	.463	.469	.488	.516	.495	.516	.381	.426	.500	.530	.581	.667	—	.643	.554	.520	6.6	1.8
14. Usage (6)	.536	.504	.491	.479	.445	.611	.392	.424	.598	.609	.599	.607	.632	—	.646	.498	6.3	1.9
15. References (6)	.516	.515	.480	.539	.478	.507	.468	.504	.526	.629	.584	.630	.603	.589	—	.614	6.6	1.1
16. Nonverbal IQ (6)	.430	.406	.402	.466	.400	.424	.379	.670	.425	.501	.417	.584	.507	.464	.549	—	103.0	13.6
M	3.7	3.4	3.9	3.8	3.7	3.4	3.9	88.8	5.1	5.4	5.8	5.8	5.4	5.1	5.7	91.4		
σ	.69	.79	1.0	.72	.74	.87	.74	14.2	.95	.86	1.2	1.2	1.5	1.5	.90	13.5		

Note. The parenthesized value following the title of each test refers to the administration year of the measure.

267–270; Kenny, 1973) was employed in modifying all cross-lagged correlations reported below. The general effect of this procedure is the attenuation of cross-lagged differences; thus, the results to follow, if inaccurate, will be biased in a conservative direction.

Results and Discussion

The intercorrelations among g_c and g_f measures in Grades 4 and 6 for both socioeconomic-status groups are presented in Table 1, along with the means and standard deviations of each test at each time period. A comparison of the mean g_c and g_f subscale scores of the two groups revealed the existence of significant differences on every measure, with scores invariably favoring the middle-socioeconomic-status sample (all t differences were significant at $p < .0001$). Comparisons of test variance between the groups indicated significantly greater variability of test scores within the middle-socioeconomic-status group on each g_c measure (all F values were significant at $p < .01$); in comparisons involving the g_f measure, however, no significant differences in test variability between the groups were encountered.

When the correlations between all measures were converted to Fisher's Z scores and were averaged and reconverted, it was found that the average correlation between the g_f measure (Lorge Thorndike Nonverbal) and the seven g_c measures was .497 in Grade 4 and .560 in Grade 6. These correlations are within the range predicted by the theory (Cattell, 1971, p. 99; Horn, 1968). The average correlation among the seven g_c measures was .580 in Grade 4 and .670 in Grade 6. The various crystallized measures apparently assess somewhat different components of the (partly) culturally determined g_c composite and, thus, intercorrelate moderately but not highly with each other (cf. Horn, 1968).

The corrected cross-lagged correlations between the g_f and g_c measures and their associated t values are presented in Table 2.[8]

TABLE 2

Corrected Cross-Lagged Correlations between Fluid Ability (G_F) and Crystallized Ability (G_C) Measures for Middle- and Low-Socioeconomic-Status Samples

Measures	Socioeconomic status			
	Middle (n = 3,994)		Lower (n = 1,501)	
	r	t	r	t
LT-NV → Vocabulary	.4844	2.182*	.4090	−.530
Vocabulary → LT-NV	.4557		.4220	
LT-NV → Reading Comprehension	.4863	2.248*	.3147	−1.50
Reading Comprehension → LT-NV	.4531		.3586	
LT-NV → Spelling	.3963	1.056	.3931	−.823
Spelling → LT-NV	.3814		.4127	
LT-NV → Capitalization	.4900	2.569*	.4476	−.767
Capitalization → LT-NV	.4539		.4676	
LT-NV → Punctuation	.4478	1.427	.4073	−.437
Punctuation → LT-NV	.4212		.4185	
LT-NV → Usage	.4651	2.240*	.4195	−.363
Usage → LT-NV	.4342		.4283	
LT-NV → References	.5355	5.051*	.4465	.717
References → LT-NV	.4660		.4284	

Note. Abbreviation: LT-NV = Lorge-Thorndike Nonverbal scale.
* $p < .05$.

These findings indicate rather strong support for the theory of g_f and g_c in the middle-socioeconomic-status sample. In every case, for example, discrepancies between cross-lagged correlations were in the direction predicted, and in five of seven comparisons, these discrepancies attained statistical significance (in the remaining instances, one comparison attained the $p < .10$ level of significance). Since the theory recognizes numerous causes of g_c in addition to g_f, the pattern of causal determination presented in Table 2 is especially noteworthy. The fact that none of the other leading theories of intelligence (e.g., Burt, 1949; Guilford, 1967; Thurstone, 1938; Vernon, 1961) predict the obtained pattern

[8] The t test employed in this investigation was based on a correction to the usual test of differences between correlations, suggested originally by Pearson and Filon (1898). In this analytic procedure, the indirect correlation between the arrays under comparison are modified by the four other relevant values (see also Peters & Van Voorhis, 1940).

of causal relations lends particular value to the data of this investigation. In attempting to weigh the merits of one of these theories against the others, researchers in the future must come to grips with this rather compelling evidence in favor of the Cattell-Horn formulation.

A final observation on the findings of this investigation concerns the results obtained among the lower-socioeconomic-status group. In this subsample, the theory finds no support whatsoever. All but one of the obtained causal relationships are in the direction opposite to that predicted, and the single confirmatory finding (that for the Knowledge and Use of References subtest) did not reach significance.[9]

Why g_f should show a strong tendency to act causally with respect to g_c among middle- but not lower-socioeconomic-status schoolchildren is not immediately obvious; certainly nothing in the theory predicts this outcome. It is possible, however, that the postulated causal mechanism operates only in the presence of a certain previously existing minimum level of g_c, with the acquisition of this minimum g_c level depending almost solely on cultural and educational exposure opportunities (see Crano, 1973). In the present research, as in previous studies, large and significant differences in test scores were found favoring the middle-socioeconomic-status group; these differences were found at both grade levels and on all measures. It is therefore conceivable that the great bulk of the lower-socioeconomic-status sample had not acquired the minimum levels of competency necessary in the seven g_c areas assessed to activate the $g_f \rightarrow g_c$ causal mechanism. The relative depression of the means and variances of the g_c measures in the lower-socioeconomic-status sample lends support to the plausibility of this speculation. This hypothesis could be further tested by comparing the patterns of causality obtained in different g_c score ranges within low- and middle-socioeconomic-status groups.

An alternative explanation can perhaps be found in the area of motivational differences. The g_f is postulated as a necessary but not sufficient cause for g_c, as is individual motivation (Horn & Cattell, 1967). If individual motivation to acquire g_c abilities is relatively uniform (and moderately high) in middle-socioeconomic-status children, the role of motivation in creating individual differences in g_c would be expected to be small relative to that of g_f. Conversely, if individual motivational levels were more variable in the lower-socioeconomic-status sample, the relative causal influence of g_f on g_c would be expected to be smaller and that of motivation larger.

This motivation-based proposition would appear to call for greater g_c variability in the lower-socioeconomic-status group, a finding opposite to that obtained in this research. It is conceivable, however, that motivational differences would affect g_c variance only after some minimal threshold achievement level is surpassed. On the basis of this reasoning, we would assume that even high levels of motivation will have little implication for achievement among children of very low g_c. In future studies, the motivational-differences hypothesis can be tested by cross-lagging measures of motivation with g_c measures in lower- and middle-socioeconomic-status groups with the prediction that motivation would exert the strongest causal influence on g_c in lower-socioeconomic-status groups of moderate-to-high academic achievement.

In summary, the present study provides strong support for the proposition, derived from the Cattell-Horn theory of intelligence, that g_f is related in a causal manner to g_c, at least in middle-socioeconomic-status children. The data from the low-socioeconomic-status sample, however, did not support this prediction: g_f apparently has little implication for g_c among children of low socioeconomic sta-

[9] The four tests excluded from the analysis show the same pattern of results. All differences are in the predicted direction for the middle-socioeconomic-status sample and in the opposite direction for the low-socioeconomic-status group. For Map Reading, Graph and Table Reading, Arithmetic Concepts, and Arithmetic Problems, t values were 2.248, 2.254, 1.105, and 1.404, respectively, in the middle-socioeconomic-status group, and −1.500, −.850, −.055, and −3.060, respectively, in the low-socio-economic-status group. However, because of possible contamination by factors other than g_c, results for these measures should be interpreted with caution, if at all.

tus (and low levels of achievement). Possible avenues for exploring this discrepancy have been proposed and will hopefully lead to a more complete understanding of the mechanisms which underly these apparent differences in the causal factors influencing academic and intellectual development.

REFERENCES

BURT, C. The structure of the mind. *British Journal of Educational Psychology*, 1949, **19,** 190.

BURT, C. The evidence for the concept of intelligence. *British Journal of Educational Psychology*, 1955, **25,** 158–177.

CAMPBELL, D. T. From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

CATTELL, R. B. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 1963, **54,** 1–22.

CATTELL, R. B. *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin, 1971.

CRANO, W. D. Causal analyses of the effects of socioeconomic status and initial intellectual endowment on patterns of cognitive develpment and academic achievement. In D. R. Green (Ed.), *The aptitude-achievement distinction*. New York: McGraw-Hill, 1973.

CRANO, W. D., & BREWER, M. B. *Principles of research in social psychology*. New York: McGraw-Hill, 1973.

CRANO, W. D., KENNY, D. A., & CAMPBELL, D. T. Does intelligence cause achievement? A cross-lagged panel analysis. *Journal of Educational Psychology*, 1972, **63,** 258–275.

GUILFORD, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1967.

HORN, J. L. The organization of abilities and the development of intelligence. *Psychological Review*, 1968, **75,** 242–252.

HORN, J. L., & CATTELL, R. B. Age differences in primary mental ability factors. *Journal of Gerontology*, 1966, **21,** 210–220. (a)

HORN, J. L., & CATTELL, R. B. Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology*, 1966, **57,** 253–276. (b)

HORN, J. L., & CATTELL, R. B. Age differences in fluid and crystallized intelligence. *Acta Psychologica*, 1967, **26,** 107–129.

KENNY, D. A. Cross-lagged and synchronous common factors in panel data. In A. S. Goldberger and O. D. Duncan (Eds.), *Structural equation models in the social sciences*. New York: Academic Press, 1973.

LINDQUIST, E. F. & HIERONYMUS, A. N. *Iowa Tests of Basic Skills: Manual for administrators, supervisors, and counselors*. Boston: Houghton Mifflin, 1964.

LORGE, I., THORNDIKE, R. H., & HAGAN, E. *Lorge-Thorndike Intelligence Tests: Technical Manual*. Boston: Houghton Mifflin, 1966.

PEARSON, K., & FILON, L. N. G. Mathematical contributions to the theory of evolution. *Transactions of the Royal Society*, Series A, 1898, **191,** 259, 262.

PETERS, C. C., & VAN VOORHIS, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.

PELZ, D. C., & ANDREWS, F. M. Detecting causal priorities in panel study data. *American Sociological Review*, 1964, **29,** 836–848.

THURSTONE, L. S. *Primary mental abilities*. Chicago: Chicago University Press, 1938.

VERNON, P. E. *The structure of human abilities*. London: Methuen, 1961.

(Received March 22, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 262-266

TRAINING IMAGERY PRODUCTION IN YOUNG CHILDREN THROUGH MOTOR INVOLVEMENT[1]

WILLIAM H. VARLEY,[2] JOEL R. LEVIN,[3] ROGER A. SEVERSON, AND PETER WOLFF

University of Wisconsin—Madison

Kindergarten and first-grade children were given a paired-associate learning task following one of five types of strategy-training procedures. In the motor-training conditions, subjects generated interactions involving pairs of toys by playing with them or by drawing pictures of them. It was found that relative to simple imagery practice, motor training facilitated the performance of kindergartners, with no differences among four motor-training variations. In the first grade, imagery practice by itself was as effective as each of the motor-training procedures. The results were discussed in terms of Piaget's theory of cognitive development and were contrasted with previously unsuccessful attempts to induce self-generated elaboration strategies in young children.

It is well established that paired-associate learning is facilitated by the addition of experimenter-provided elaborations (Rohwer, 1967). However, this is not always true when young children (typically below eight years of age) are asked to generate such elaborations themselves (Levin, 1972; Rohwer, 1972; but see McCabe, Levin, & Wolff, in press). Given that young children are unsuccessful in generating sentence and imagery elaboration strategies on request, the possibility exists that with systematic instruction or training they may be taught to do so. However, initial efforts to train mnemonic elaboration in young children have been disappointing (Rohwer & Ammon, 1971; Rohwer, Ammon, & Levin, 1971), leading Rohwer (1972) to conclude that children must attain a certain age or maturational level before they can be taught to use elaboration strategies effectively.

Considering the above failures, such a conclusion is tempting. On the other hand, it has been recently demonstrated that (supposedly) "pre-imagery" children can be induced to generate facilitative imagery elaborations through concurrent motor involvement (Wolff & Levin, 1972). In that experiment, it was found that although six- and seven-year olds did not benefit from instructions to imagine an interaction between pairs of toys, when they were permitted to play with toys concurrently, their performance improved dramatically (even though they were not allowed to see the manipulations they generated). This result was interpreted as being consistent with Piaget's theory of cognitive development, in which it is assumed that the preoperational child cannot produce dynamic visual representations (internally) without motor activity involving the events to be represented (Piaget, 1962).

Danner and Taylor (1973) were also able to induce imagery in first graders by giving them pretraining in drawing pictures of separate objects interacting. Although these

[1] The research reported here was done as part of the first author's doctoral dissertation, with support from the Wisconsin Research and Development Center for Cognitive Learning at the University of Wisconsin, Madison. We are grateful to Billie Albrecht for typing the final draft of the paper.

[2] Now at Dartmouth-Hitchcock Mental Health Center, Hanover, New Hampshire.

[3] Requests for reprints should be sent to Joel R. Levin, Wisconsin Research and Development Center, 1025 West Johnson Street, Madison, Wisconsin 53706.

investigators did not interpret their findings in terms of motor involvement, Piaget and Inhelder (1971) would view playing and drawing activities similarly, with each resulting in internalized imitation (visual imagery). Thus, the important distinction between the Danner-Taylor and Wolff-Levin experiments is that in the former, the motor activity (i.e., the pretraining) was temporally removed from the criterion performance (and involved different items), whereas in the latter, the motor activity and criterion performance were concurrent.

The purpose of the present study was twofold: (*a*) to investigate differences in performance of younger and older children within the imagery-transition stage and (*b*) to extend the motor–imagery separation results to other types of training procedures.

Method

Subjects

Eighty children apiece from the kindergarten and first grade of a small, middle-class community in southern Wisconsin served as the subjects.[4] All testing was done during the spring of 1972. At that time, the median age of the kindergartners was 6 years, 1 month, and that of the first graders was 7 years, 1 month. The subjects were taken in turn from their classrooms and randomly assigned (in equal numbers) to one of five treatment conditions at the time they entered the experimental room.

Design and Materials

Following Wolff and Levin (1972), a paired-associate task was constructed with small children's toys comprising the stimulus materials. Fifteen pairs were created for the learning task, with an additional eight pairs used for practice and examples during training. All pairs could be easily labeled by kindergarten and first-grade children, as determined from an earlier pilot study. Each pair (e.g., cowboy–car) was formed randomly, subject to the constraint that a plausible interaction existed for the two paired toys.

Prior to the learning task, the subjects were given one of five types of strategy training. In the imagery control condition, the experimenter demonstrated a predetermined interaction for each of the first four training pairs, after which the subject was given instructions to imagine an interaction for each of the four remaining pairs. This nonmotor–imagery condition was incorporated as a baseline for evaluating the effects of four motor–imagery training procedures in each grade, as follows: In two conditions, the subjects were instructed to generate interactions for each of the eight training pairs, either by playing with the toys (repeated play) or by drawing a picture of them in interaction (repeated draw). These conditions approximated those used by Wolff and Levin (1972) and Danner and Taylor (1973), respectively. Finally, in another two conditions, the subjects were given playing (or drawing) practice on the first four training pairs, followed by a delayed play (or draw) instruction for the remaining four pairs in which the subjects were required to indicate that they had generated an imaginal interaction *before* they were permitted to execute it through playing with (or drawing) the toys. It was assumed that these latter two conditions—faded play and faded draw—would produce better subsequent performance than their repeated counterparts (especially for kindergartners), since the subjects would be receiving practice in generating imagery apart from motor activity.

Procedure

In order to monitor the elaboration of the subjects during training, interactions created by the subject or the experimenter were briefly described by the experimenter without labeling the toys (e.g., "Look, it's chasing that."). The subject's verbalizations regarding his interactions were suppressed by the experimenter during training so that facilitative effects of subject-generated verbalization would not cloud an imagery interpretation of the training effects. It was difficult to equate the amount of time across training conditions. Altogether, the subject spent about 25 minutes with the experimenter in the imagery control, repeated play, and faded play conditions, and about 35 minutes in the repeated draw and faded draw conditions.

The learning task itself was presented by way of an incidental-learning format because it was thought to be less reactive in this kind of training experiment, and since no differences in performing the task as a function of intention to learn have been found with subjects of this age (cf. Wolff, Levin, & Longobardi, in press). The subject's introduction to the learning task was the same across all conditions, but the "set" given to each subject was based on the training he had received.[5]

All subjects were tested individually. Following the 8 training items, the 15 learning task pairs were presented one at a time at a 10-second rate (timed by the experimenter). After the last pair

[4] Thanks are due to the staff and students of the Sauk Prairie Public Schools in Sauk Prairie, Wisconsin.

[5] The instructions given to the children within the various conditions, and a list of the stimulus materials employed, are available from the second author on request.

was presented, the 15 response toys were displayed, and for each stimulus toy (presented in a different random order than the study items), the subject was required to pick up the response toy with which it was initially paired. The subjects were given 7 seconds to respond to each stimulus toy. Each of the subject's selections was replaced in the array before the next stimulus toy was presented.

Results

Learning was defined as the total number of correct responses (out of 15) during the recognition test. The mean number of correct responses, by condition and grade, is presented in Figure 1. In the analysis of variance, comparisons were nested within grades so that statements about the effectiveness of the different conditions could be made separately for each grade. To do this, each motor-training condition was compared with the imagery control condition via Dunnett's test (one-tailed, $\alpha = .05$). In addition, the factorial combination of the nature (play or draw) and extent (repeated or faded) of motor involvement yielded three orthogonal comparisons among training procedures within each grade.

Within the kindergarten sample, Dunnett's comparisons revealed that each motor-training condition produced significantly better learning than did the imagery control condition. However, within the first-grade sample, mean performance in the imagery control condition was as good as in each of the motor-training conditions. No significant differences among the four motor-training conditions were detected in either grade (all $ps > .05$).

A post hoc breakdown of these effects yielded further interesting information. When the data were analyzed separately for boys and girls, it was found that the comparability of the imagery control and motor-training conditions in the first grade was true only for girls ($|t| < 1$); for boys, however, significant differences in favor of motor training still existed ($t = 1.86$, $df = 29$, $p < .05$, one-tailed) as they had for both sexes in the kindergarten sample (girls, $t = 2.96$, $df = 32$, $p < .01$; boys, $t = 2.57$, $df = 38$, $p < .025$). Despite these different effects for boys and girls, the previous finding of no significant differences among the four motor-training conditions did not change when investigated for each sex group separately.

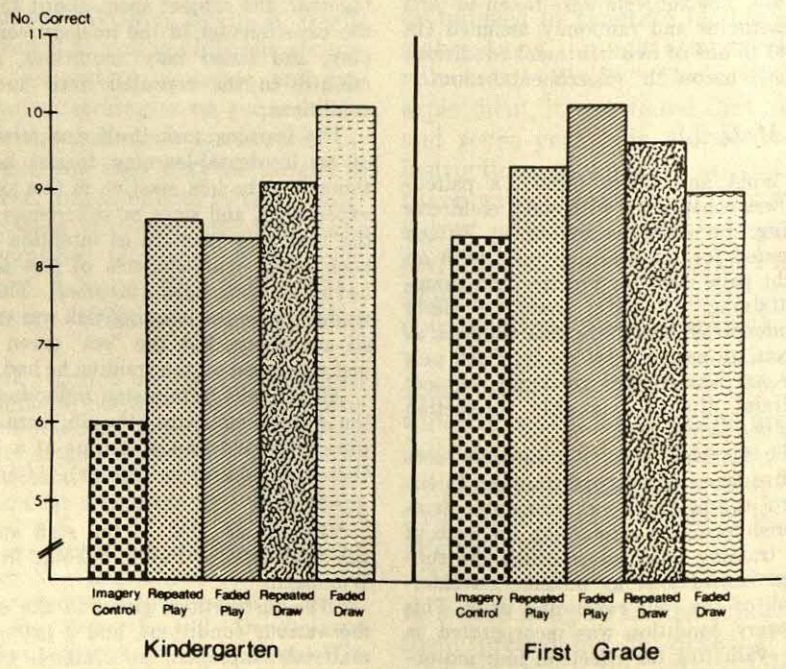


Figure 1. Mean number of correct responses by grade and experimental condition.

Discussion

The results of the present study indicated that for kindergarten children, each motor-training condition resulted in significantly better learning than the nonmotor–imagery control condition. However, the effect due to motor training was not observed in the first-grade sample where motor and nonmotor training resulted in nonsignificant learning differences. It might therefore be inferred that between the ages of six and seven the young child is becoming increasingly adept at generating dynamic representations in the absence of concurrent motor involvement or without immediate prior training. This age range may well be adjusted depending on the sex of the child (with the present data suggesting that the ability develops earlier in girls than in boys) as well as on the particular sociocultural characteristics of the population being considered.

The finding that playing and drawing training did not differ statistically is consistent with Piaget and Inhelder's (1971) view that both represent vehicles for visually internalizing external stimuli. The failure of the faded training conditions to produce learning superior to the repeated conditions was unexpected. However, it was noted by the experimenter that the subjects in the repeated conditions typically hesitated before either playing with or drawing the toys for each training pair. This observation suggests that the subjects in these conditions were thinking up (imagining?) interactions *before* they executed the motor activity, even though they were not explicitly instructed to do so as in the faded conditions. Such an observation, of course, also hints at additional research (utilizing response latencies, delayed visual feedback, motor blockage, and other variables) to specify more precisely the temporal antecedents of imagery production in young children.

The results of this study have implications for future efforts to train imagery elaboration in children of this age. They strongly support the assumption that procedures designed to train cognitive skills must take into account the information-processing skills of the young child as postulated by developmental learning theories. The present training procedures, based on Piaget's theory of cognitive development, indicate that motor involvement should be a dominant feature of imagery training in young children, an approach not fully exploited by previously unsuccessful training studies (Rohwer & Ammon, 1971; Rohwer, Ammon, & Levin, 1971).

It should be noted that the learning materials of the present study consisted of toys, which may be regarded as being more concrete (in Paivio's, 1971, terms) than the pictures and aurally presented words of Rohwer's (1967) earlier work and which may be assumed to evoke dynamic images more readily (Levin, 1972). However, perhaps one of the most exciting results here was the marked facilitative effects of the drawing-training conditions, substantiating those of Danner and Taylor (1973). A worthwhile direction for future training studies with young children would be to apply drawing training to the learning of less concrete materials. In this way, children could be taught to elaborate pictorial or aural materials while honoring the need for motor involvement in the training process. The fact that the playing and drawing conditions resulted in equal degrees of facilitation here strengthens this conclusion.

REFERENCES

Danner, F. W., & Taylor, A. M. Integrated pictures and relational imagery training in children's learning. *Journal of Experimental Child Psychology*, 1973, **16**, 47–54.

Levin, J. R. When is a picture worth a thousand words? In, *Issues in imagery and learning: Four papers*. (Theoretical Paper 36) Madison: Wisconsin Research and Development Center for Cognitive Learning, February 1972.

McCabe, A. E., Levin, J. R., & Wolff, P. The role of overt activity in children's sentence production. *Journal of Experimental Child Psychology*, in press.

Paivio, A. *Imagery and verbal processes*. New York: Holt & Co., 1971.

Piaget, J. *Play, dreams, and imitation in childhood*. New York: Norton, 1962.

Piaget, J., & Inhelder, B. *Mental imagery in the child*. New York: Basic Books, 1971.

Rohwer, W. D., Jr. *Social class differences in the role of linguistic structures in paired-associate learning: Elaboration and learning proficiency*.

(Final report on basic research Project 5-0605) Wash., D.C.: U.S. Office of Education, 1967.

ROHWER, W. D., JR. Decisive research: A means for answering fundamental questions about instruction. *Educational Researcher,* 1972, **1**(7), 5–11.

ROHWER, W. D., JR., & AMMON, M. S. Elaboration training and paired-associate learning efficiency in children. *Journal of Educational Psychology,* 1971, **62,** 376–383.

ROHWER, W. D., JR., AMMON, M. S., & LEVIN, J. R. Learning efficiency and elaboration training among four- and five-year old children. In W. D. Rohwer, Jr., & P. R. Ammon, The assessment and improvement of learning and language skills in four- and five-year old culturally disadvantaged children. Final report, Office of Economic Opportunity, Contract No. B99-4776, June 1971.

WOLFF, P., & LEVIN, J. R. The role of overt activity in children's imagery production. *Child Development,* 1972, **43,** 537–547.

WOLFF, P., LEVIN, J. R., & LONGOBARDI, E. T. Activity and children's learning. *Child Development,* in press.

(Received March 22, 1973)

ANNOUNCEMENT

In addition to the members of the Advisory Board, the *Journal of Educational Psychology* asks other psychologists to aid in the review of manuscripts. In this way we have been able to achieve a broader review base. The Editors wish to thank the following people for their generous assistance to the Journal.

MOSHE ANISFELD, *Yeshiva University*
ALBERT BEATON, *Educational Testing Service*
BERNARD BEER, *Squibb & Sons, Inc., New Brunswick*
GORDON H. BOWER, *Stanford University*
COURTNEY B. CAZDEN, *Harvard University*
EDWARD A. CHITTENDEN, *Educational Testing Service*
JOEL B. COHEN, *National Analysts, Inc., Philadelphia*
EDMUND B. COLEMAN, *University of Texas at El Paso*
SHEL FELDMAN, *Swarthmore College*
JEAN E. FERSON, *Temple University*
NORMAN FREDERICKSEN, *Educational Testing Service*
KENNETH GOODMAN, *Wayne State University*
CURTIS D. HARDYCK, *University of California, Berkeley*
RODNEY HARTNETT, *Educational Testing Service*
VIVIAN M. HORNER, *Yeshiva University*
DEAN JAMISON, *Educational Testing Service*
JEROME KAGAN, *Harvard University*
PAUL A. KOLERS, *University of Toronto*
LUIS M. LAOSA, *University of California, Los Angeles*
RICHARD LIGHT, *Harvard University*
ELEANOR E. MACCOBY, *Stanford University*
DENNIS MOLFESE, *Southern Illinois University*
CAROL MOORE, *Educational Testing Service*
SEYMOUR PAPERT, *Massachusetts Institute of Technology*
DAVID POTTER, *Educational Testing Service*
THOMAS QUIRK, *Principia College*
JOSEPH M. SCANDURA, *University of Pennsylvania*
FRED STRASSBURGER, *American Psychological Association*
ROBERT THORNDIKE, *Teachers College, Columbia University*
BRUCE W. TUCKMAN, *Rutgers—The State University*
ESTER WESTERVELT, *Simmons College*
KENNETH WILSON, *Educational Testing Service*

Journal of Educational Psychology
1974, Vol. 66, No. 2, 267–276

SOCIAL-EMOTIONAL, COGNITIVE, AND DEMOGRAPHIC DETERMINANTS OF POOR SCHOOL ACHIEVEMENT: IMPLICATIONS FOR A STRATEGY OF INTERVENTION[1]

MARTIN KOHN[2] AND BERNICE L. ROSMAN

William Alanson White Institute of Psychiatry, Psychoanalysis and Psychology, New York, New York

The effect of three classes of variables (preschool cognitive functioning, preschool social-emotional functioning and background-demographic variables) on early elementary school achievement was examined. Two hundred and nine black and white boys from lower- and middle-class backgrounds were evaluated during the preschool period and received achievement tests during the second year of elementary school. Each of the three classes of variables accounted for a significant proportion of the variance of the criterion measures. When the classes of variables were examined by means of hierarchical regression technique, the social-emotional and cognitive variables yielded the most information for programs of psychological intervention. Intervention directed at the social-emotional components of cognitive performance was discussed.

Social scientists have been increasingly concerned with the antecedents of school success. Attempts at early cognitive remediation, such as some Head Start and Home Study programs, have been based on the assumption that disadvantaged home environments fail to provide the early learning experiences needed to cope with the intellectual demands of school. These intervention programs have focused primarily on the remediation of *cognitive handicaps* and have paid little attention to the relationship between *social-emotional functioning* and learning difficulties and underachievement.

Wechsler (1971) has suggested that intellectual achievement is "dependent, to varying degrees, upon a variety of determinants which are more of the nature of connative or personality traits rather than of cognitive abilities [p. 51]." These observations are strongly confirmed in the clinical literature on emotional disturbance and learning difficulties (Pearson, 1952, 1954); systematic evidence is sparse.

Kohn and Rosman (1972a) have developed a two-factor model of the child's social-emotional functioning in the preschool setting as follows: Factor I—Interest-Participation versus Apathy-Withdrawal; Factor II—Cooperation-Compliance versus Anger-Defiance. They have established that these dimensions (*a*) assess relatively enduring attributes of children across settings and over time (Kohn & Rosman, 1972a, 1973b), (*b*) are valid measures of emotional health–disturbance (Kohn & Rosman, 1973c), and (*c*) are relevant to preschool cognitive functioning and to later school achievement (Kohn & Rosman, 1972b, 1973a). Apathy-Withdrawal, but not Anger-Defiance was predictive of poor achievement (Kohn & Rosman, 1972b). More

[1] This research was carried out under Grant 16944 from the National Institute of Mental Health to the first author.

The authors gratefully acknowledge the excellent cooperation extended by the Division of Day Care of the New York City Department of Social Services, the Day Care Council of New York, Inc., and the New York City Board of Education. The authors are particularly indebted to their statistical consultant, Jacob Cohen, for the application of his innovative statistical technique to the problems and issues set forth in this study.

[2] Requests for reprints should be sent to Martin Kohn, William Alanson White Institute, 20 West 74th Street, New York, New York 10023.

recently, a third factor dimension of social-emotional functioning (Task Orientation) was found to be highly correlated with cognitive functioning at the preschool level (Kohn & Rosman, 1973a).

The purpose of the present study was two-fold:

1. To determine the extent to which early elementary school achievement is a function of each of three classes of variables: preschool social-emotional functioning, preschool cognitive functioning, and background-demographic variables. It was hypothesized that low achievement is predicted by social-emotional disturbance, by cognitive impairment, and by low status on a series of background-demographic variables related to social advantage–disadvantage. Since the cognitive variables were in the same domain as the criterion measures, they were expected to be the best predictors.

2. To generate information expected to be useful in remediation and intervention programs by examining the joint effect of the three classes of variables on second-grade school achievement. Basic to the study was the assumption that deficits and disturbances in preschool social-emotional and cognitive functioning would be more remediable, at least by psychologists, than disadvantaged status as reflected in the background variables. The following questions were posed:

1. To what extent do the two classes of psychological variables (i.e., the social-emotional and cognitive variables) predict over and above their separate effects?

2. In what sequence do these two classes of variables yield the maximum amount of information, that is, when social-emotional variables are examined over and above the variance accounted for by the cognitive variables or vice versa?

3. To what extent are the background-demographic variables predictive of second-grade school achievement after the psychological measures have been partialed out?

The individual variables within each class and their relevant hypotheses appear in the next section.

Method

Subjects

Subjects were 209 second-grade boys selected two years earlier when they were enrolled in New York City public school kindergartens.[3] Kindergarten is referred to as "preschool" simply to indicate that it is the period prior to the onset of formal education and that emotional impairment during this period is unlikely to be due to academic failure. Originally, 217 children were included in the study; vigorous follow-up efforts resulted in the location of 96.4% of the sample in public, private, and parochial schools in the metropolitan and suburban New York area as well as in other states.

Subjects had been selected, in approximately equal numbers, from three social class levels based on the Hollingshead Two-Factor Index of Social Position[4]; ranked from low to high, these levels are Class V, IV, and III plus (III or better). Half of the sample was white and half black; black and white subjects were distributed almost equally across social class levels. (Further details concerning selection and additional background characteristics of the sample during the preschool period appear in Kohn and Rosman, 1973a.)

The predictor variables (social-emotional functioning, cognitive functioning, and background-demographic data) were collected during the preschool period while subjects were attending kindergarten. Their mean age at the time was 68 months (5 years 8 months). The criterion variables (achievement measures) were collected when the children were in second grade; mean age at the time was 88 months (7 years 4 months).

Measures of Social-Emotional Functioning

Instruments. Social-emotional functioning was assessed by means of three instruments: Kohn Social Competence Scale, Kohn Problem Checklist (Kohn & Rosman 1972a), and Schaefer Classroom Behavior Inventory.[5] The two Kohn scales each measure two major dimensions of social-emotional functioning: (*a*) Factor I—Interest-Participation versus Apathy-Withdrawal and (*b*) Factor II—Cooperation-Compliance versus Anger-Defiance.

Factor I indicates the extent to which the child displays interest and curiosity, utilizes the opportunities of the classroom, and interacts positively

---

[3] Included in the original sample were 70 children from New York City day care centers.

[4] A. B. Hollingshead. Two-factor index of social position. Unpublished manuscript, 1957. (Available from A. B. Hollingshead, Department of Sociology, Yale University, New Haven, Connecticut 06510.)

[5] E. S. Schaefer and M. R. Aaronson. Classroom behavior inventory: Preschool to primary. Unpublished manuscript, 1966. (Available from E. S. Schaefer, Department of Maternal Health and Child Care, Duke University, Durham, North Carolina 27701.)

with his peers or, on the other hand, to what extent he withdraws and displays little interest in his peers or in kindergarten activities. High ratings on Apathy-Withdrawal had been associated with poor cognitive functioning at the preschool level (Kohn & Rosman, 1973a); in the present study it was hypothesized that children high on Apathy-Withdrawal would continue to perform more poorly in elementary school than those high on Interest-Participation.

Factor II indicates the extent to which the child can accept the routines of a classroom and comply with teachers' requests or suggestions or, on the other hand, to what extent he is defiant and creates disturbances which disrupt normal classroom routines. Only minimally related to preschool cognitive functioning, this factor was expected to be nonpredictive of the second-grade achievement criteria.

The Schaefer Inventory yields three major dimensions: (*a*) Factor I—Extroversion versus Introversion, (*b*) Factor II—Adjustment versus Maladjustment, and (*c*) Factor III—High versus Low Task Orientation.

In previous work (Kohn & Rosman, 1972a), the Schaefer Factor I and II dimensions were found to be congruent and substantially correlated with the Kohn Factors I and II. Schaefer Factor III, Task Orientation, which indicates the extent to which the child is able to concentrate, become absorbed, and persevere in activities or, on the other hand, to what extent he is hyperactive and distractible, has a low attention span, and shows low motivation to "stick to" tasks, has no equivalent on the Kohn scales. Substantially related to preschool cognitive functioning (Kohn & Rosman, 1973a), this variable was expected to be one of the better predictors of second-grade achievement.

Procedure. Each child was rated once by his kindergarten teacher on each of the three instruments. To achieve greater stability of the measures, the corresponding scores from the three instruments were converted to standard scores and pooled. The resulting pooled Factor I and II scores were given the Kohn factor designations for purposes of convenience and consistency. Schaefer Factor III was retained without pooling since there is no Kohn factor equivalent. This dimension was substantially correlated with Pooled Factor II ($r = .68$) in line with previous findings (Kohn & Rosman, 1972a[6]). Interrater reliabilities of the pooled dimensions had been adequate (*rs* ranging from .76 to .90) in an earlier study (Kohn & Rosman, 1972a). On each of the three instruments used, a high score indicated disturbance.

[6] E. S. Schaefer, L. F. Droppleman, and A. F. Kalverboer. Development of a classroom behavior checklist and factor analyses of children's school behavior in the U.S. and the Netherlands. Unpublished manuscript, 1965. (Available from E. S. Schaefer, Department of Maternal Health and Child Care, Duke University, Durham, North Carolina 27701.)

Measures of Cognitive Functioning

Instruments. Five measures made up the class of cognitive predictors. The Stanford-Binet Intelligence Scale (Form L-M) is a widely used, well-standardized measure of general intelligence that is highly predictive of later school achievement and learning. The four other measures were derived from a factor analysis of a large number of cognitive tests tapping various areas of cognitive functioning; the specific tests have been described in Kohn & Rosman (1973a).

1. Factor I—Visual Cognition. This factor subsumes a group of tasks demanding visual attention, analysis of visual material, and processing of visual cues. Nonverbal in presentation and performance, the tasks are composed of simple and concrete visual stimuli that minimize previous learning but require active mental processes such as classifying and reasoning. This factor was expected to show low correlation with school achievement, particularly Word Knowledge and Reading.

2. Factor II—Verbal Expressivity reflects the child's readiness to speak, his verbal productivity and the coherence and clarity of his verbal communications; it was expected to account for only a trivial amount of variance in the achievement measures.

3. Factor III—Motor Control indicates the extent to which the child is able to follow directions in carrying out a task and may also reflect his ability to channel his energies into task-appropriate activities. Children who scored low on this factor were expected to perform more poorly than those who scored high.

4. Factor IV—Verbal Cognition subsumes a group of tasks commonly assumed to measure "verbal ability." Major components of the factor include passive and active vocabulary, verbal concepts, auditory attention, discrimination, and memory. This factor was expected to be most relevant to school achievement, particularly Reading and Word Knowledge.

Because factor analysis of preschool intelligence tests have revealed factorial dimensions similar to the four cognitive factors of the present study (Stott & Ball, 1965), it was expected that these four factors would, jointly, predict as well as the Stanford-Binet. The Stanford-Binet itself was expected to be a better predictor than any one of the four factors.

Procedure. The children were tested individually in five half-hour sessions on the series of cognitive tasks by four testers trained at the master's level in school psychology. The testers were female, white, and in their twenties; no evidence was found of any systematic bias or other differences attributable to the different testers.

Background-Demographic Measures

Home interviews with mothers yielded data on five background variables previously hypothesized to be relevant to school achievement (Kohn & Rosman, 1973a).

1. Social class as measured by the Hollingshead index; occupation and education of head of household were scaled, differentially weighted, and combined; the resulting scores formed a continuum ranging from 11 (low) to 77 (high).

2. Race (0 = black, 1 = white).

3. Welfare status (0 = family receives welfare assistance, 1 = no assistance received).

4. Family size (number of children in nuclear family).

5. Family intactness (2 = child lives with both natural parents, 1 = other).

Achievement Measures

Three parts of the Metropolitan Achievement Test (Form B) were administered to assess second-grade achievement: Word Knowledge, Reading, and Arithmetic Concepts and Skills. All of the subjects were tested, individually or in pairs, on the Primary I battery; children receiving a perfect score on one or more parts of the test were further tested on Primary II. Raw scores on the three subtests were converted to grade equivalents on the basis of the Metropolitan Achievement Test norms. Testing was carried out by three college student research assistants trained to administer the instruments. These students were in their twenties, female, and white. Seven children who lived outside of the metropolitan area were tested by their own second-grade teachers.

Data Analysis

The data were analyzed by means of correlational and hierarchical multiple regression techniques. Cohen (1968) had demonstrated the equivalence of multiple and partial regression on the one hand and analysis of variance and analysis of covariance on the other hand. In the hierarchical multiple regression technique, the independent variables are arrayed in sets, and the multiple correlation coefficients are determined as successive sets are added to the analysis. The increment in multiple correlation squared obtained by adding a new set to the analysis may be interpreted as the proportion of variance accounted for by the new set after variance associated with the preceding sets has been partialed out. An F test of the significance of the increment is easily carried out.

Sets that are significant are further examined to determine which of the variables within the set contribute to the significance of the set. Partialed out from a specific variable in a given set are the variables in the preceding sets (if any) as well as the remaining variables within the same set. The cumulative contribution of all the sets gives the final multiple correlation squared with the dependent variables.

Two series of analyses were carried out. The first examined the amount of variance accounted for by each of the three classes of independent variables and tested the hypotheses about the individual variable in each class. The second series examined the effects obtained by combining classes of independent variables. The following were determined: first, the amount of variance accounted for by social-emotional variables after the cognitive variables had been partialed out and second, the amount of variance accounted for by the cognitive variables after the social-emotional variables had been partialed out. These analyses also yielded information about the total amount of variance accounted for by the psychological variables. Next, the amount of variance accounted for by the background-demographic variables after the two classes of psychological variables had been partialed out was determined. This final step also yielded information about the proportion of variance accounted for uniquely by the background variables.

Results

By Class of Variables

The extent to which the three classes of variables predicted the achievement criteria is shown in Table 1. The percentage of variance in the dependent measures accounted for by each class of variables is listed under the heading $R^2\%$ (i.e., $R^2 \times 100$); the zero-order correlations between the individual variables and the criteria appear in the column headed r_o; the partial correlations (when other variables within the same set are partialed out) are given in the column headed r_p.[7]

Social-emotional variables. The three social-emotional variables together account for 16%–22% of the variance in the achievement measures. Individually, each shows significant zero-order correlation to the criteria. When the other two social-emotional variables are partialed out from each other so that the unique contribution of each can be assessed, Factors I and III, in line with expectations, show significant r_p with all three achievement measures. The findings support the hypothesis that low achievement is predicted uniquely by each

[7] Age at time of preschool testing was partialed out prior to the cognitive variables as a statistical control; this linearly age adjusted the cognitive measures. Age at time of second-grade testing was partialed out prior to the background variables; this variable was irrelevant for the kindergarten measures but the statistical control was necessary to remove spurious background age of testing effects.

TABLE 1

CORRELATION OF CLASSES OF VARIABLES AND OF VARIABLES WITHIN CLASSES TO MEASURES OF SECOND-GRADE ACHIEVEMENT

Independent variables	*df*	Achievement criteria								
		Word Knowledge			Reading			Arithmetic		
		R^2%	r_o	r_p	R^2%	r_o	r_p	R^2%	r_o	r_p
Social-emotional variables	3/205	15.5***			21.7***			16.1***		
I Apathy–Interest			−.25**	−.21**		−.21**	−.15*		−.24**	−.20**
II Anger–Cooperation			−.16*	.12		−.23**	.13		−.18*	.11
III Task Orientation			−.33**	−.30**		−.44**	−.39**		−.35**	−.31**
Cognitive variables[a]	5/202	41.4***			41.7***			40.7***		
Cognitive factors	4/203	(37.1***)			(32.2***)			(34.5***)		
I Visual Cognition			.56**	.36**		.53**	.34**		.53**	.30**
II Expressivity			.21**	.04		.16*	−.02		.28**	.12
III Motor Control			.20**	−.06		.21**	−.02		.25**	.00
IV Verbal Cognition			.52**	.30**		.50**	.28**		.52**	.28**
Stanford-Binet	1/202	(4.3***)	.59**	.26**	(8.5***)	.61**	.36**	(6.2)	.61**	.31**
Background-demographic variables[b]	5/202	19.3***			21.7***			22.5***		
Social class			.34**	.26**		.36**	.28**		.30**	.18**
Welfare			.27**	.11		.28**	.09		.37**	.24**
Family size			−.22**	−.14*		−.23**	−.17*		−.14*	−.03
Family intactness			.11	.00		.16*	.05		.22*	.05
Race			.20**	.20**		.23**	.23**		.27**	.24**

Note. Abbreviations: r_o = zero-order correlation; r_p = partial correlation with all other variables within the class partialed out. Multiple correlations squared given in parentheses for cognitive factors and the Stanford-Binet are subtotals of multiple correlations squared for the whole class of cognitive variables.

[a] Variance due to preschool test age partialed out; variance not included (df = 1).

[b] Second-grade test age partialed out; variance not included (df = 1).

* $p \leq .05$.

** $p \leq .01$.

*** $p \leq .001$.

of these two dimensions of preschool social-emotional disturbance.

Cognitive variables. The cognitive variables together account for about 41% of the variance in each of the dependent variables. The Stanford-Binet, contrary to expectations, contributes a small but significant increment over and above the effects of the cognitive factors.

In the zero-order correlations, the Stanford-Binet is the best single predictor, but Factors I and IV, surprisingly, predict almost as well. It is particularly surprising that Cognitive Factor I, composed of nonverbal tasks seemingly having little in common with achievement tests, predicts almost as well as Cognitive Factor IV and the Stanford-Binet.

When the cognitive factors are partialed from each other, only Visual Cognition (I) and Verbal Cognition (IV) remain significant; for remedial and predictive purposes, then, these are the important dimensions. The findings support the hypothesis that low school achievement is predicted by preschool cognitive level.

Background-demographic variables. Taken together, the background variables account for 19%–22% of the variance in the criterion measures. All variables show significant zero-order correlations with each of the achievement measures except for one correlation, namely, that of Family Intactness to Word Knowledge.

After the variables are partialed out from each other, higher achievement is, as hypothesized, associated with higher social class. At the zero-order level, white children performed better than black children. This difference was expected to diminish sharply after partialing; comparison of the r_os and the r_ps shows little difference.

Overall, low status on the social advantage–disadvantage continuum is related to low school achievement.

Summary. The social-emotional variables account for 16%–22% of the variance of the criterion measures, the cognitive variables for about 41%, and background-demographic variables for 19%–22% of the variance.

While it is not surprising that the cognitive variables predict twice as much variance as the background-demographic and the social-emotional variables, it is interesting that the social-emotional variables predict equally as well as the background-demographic variables.

Joint Effect of Psychological Variables

The extent to which the social-emotional and cognitive variables, taken jointly, predict the achievement criteria may be seen in Table 2. In Sequence 1, the social-emotional variables contribute very little over and above the variance accounted for by cognitive variables. Preschool social-emotional functioning and background-demographic variables have been shown to account for major proportions of preschool cognitive functioning (Kohn & Rosman, 1973a). It seems reasonable therefore to assume that the effect of social-emotional functioning on cognitive functioning has already been taken into account when preschool cognitive functioning is related to elementary school achievement.

Sequence 2, in which the social-emotional variables are entered first, yields the maximum amount of differentiated information. The four cognitive factors contribute a significant increment ranging from 17% to 23% over and above the social-emotional variables; Cognitive Factors I and IV account for unique variance over and above the other cognitive variables as well as the preceding set of social-emotional variables. The Stanford-Binet adds an additional 3%–6% to the predictions of second-grade achievement.

Why stress the social-emotional variables as predictors of school achievement when these factors are deeply embedded in the preschool cognitive variables, themselves the strongest predictors of academic success? The data in Table 2 have important implications for remediation; intervention must differentiate between the child whose poor achievement stems from cognitive deficits and the child whose learning problems are largely accounted for by social-emotional disturbance.

The final lines of Table 2 show the R^2% and the multiple correlations, as well as the

TABLE 2

JOINT EFFECTS OF THE PSYCHOLOGICAL VARIABLES (COGNITIVE AND SOCIAL-EMOTIONAL FUNCTIONING)

Independent variables	*df*	Achievement criteria					
		Word Knowledge		Reading		Arithmetic	
		R^2%	r_p	R^2%	r_p	R^2%	r_p
Sequence 1: Social-emotional variables over and above cognitive variables							
Cognitive variables[a]	5/202	41.4***		41.7***		40.7***	
Social-emotional variables	3/205	.6		3.3**		.9	
I Apathy–Interest			−.02		.04		.01
II Anger–Cooperation			.07		.09		.06
III Task Orientation			−.10		−.22**		−.11
Sequence 2: Cognitive variables over and above social-emotional variables							
Social–emotional variables	3/205	15.5***		21.7***		16.1***	
Cognitive variables[a]	5/199	26.5***		23.3***		25.5***	
Cognitive factors	4/200	(23.0***)		(16.8***)		(20.4***)	
I Visual Cognition			.33***		.30***		.27***
II Expressivity			.03		−.01		.12
III Motor Control			−.06		−.04		.00
IV Verbal Cognition			.26***		.23***		.24***
Stanford-Binet	1/199	(3.5***)		(6.5***)		(5.1***)	
Joint effects of social-emotional and cognitive variables							
R^2		42.0***		45.0***		41.6***	
Multiple regression		.65		.68		.68	
Corrected R^2		40.2		43.6		39.7	
Corrected multiple regression		.63		.66		.63	

Note. Abbreviation: r_p = partial correlation with all other variables within the set and in the preceding sets partialed out. Multiple correlations squared given in parentheses for cognitive factors and the Stanford-Binet are subtotals of multiple correlations squared for the whole class of cognitive variables.

[a] Variance due to preschool test age partialed out; variance not included ($df = 1$).

* $p \leq .05$.
** $p \leq .01$.
*** $p \leq .001$.

R^2% and the multiple correlation values corrected for shrinkage (McNemar, 1962), based on all predictors. The social-emotional variables and the cognitive variables together account for 40% to 44% (corrected) of the total variance, and the corresponding multiple correlation values are between .63 and .66.

Joint Effect of Psychological and Background Variables

Table 3 shows that the class of background-demographic variables adds a significant but relatively small proportion of variance, over and above the psychological variables, to the predictions of Reading and Arithmetic but not to the prediction of Word Knowledge. Since background-demographic variables account for a major proportion of preschool cognitive functioning (Kohn & Rosman, 1973a), it seems reasonable to assume that their effect has already been taken into account when preschool cognitive functioning is related to elementary school achievement.

The three significant partial correlations between the achievement measures and social class and welfare status suggest either (*a*) a cumulative deficit for children from deprived backgrounds, as suggested by

TABLE 3

JOINT EFFECTS OF PSYCHOLOGICAL AND BACKGROUND-DEMOGRAPHIC VARIABLES

Independent variables	*df*	Achievement criteria					
		Word Knowledge		Reading		Arithmetic	
		$R^2\%$	r_p	$R^2\%$	r_p	$R^2\%$	r_p
Preschool psychological variables[a]	8/199	42.0***		45.0***		41.6***	
Background-demographic variables[b]	5/193	2.5		4.9**		5.2**	
Social class			.18**		.23**		.11
Welfare status			.02		.03		.20**
Family size			−.06		−.11		.08
Family intactness			−.01		.02		.02
Race			.02		.06		.12
Total R^2		44.5***		49.9***		46.8***	
Multiple regression		.67		.71		.68	
Corrected R^2		40.6		46.6		43.1	
Corrected multiple regression		.64		.68		.66	

Note. Abbreviation: r_p = partial correlation with all other variables within the set and in the preceding sets partialed out.

[a] Social-emotional plus cognitive functioning. Variance due to preschool test age partialed out; variance not included ($df = 1$).

[b] Variance due to second-grade test age partialed out; variance not included ($df = 1$).

* $p \leq .05$.
** $p \leq .01$.
*** $p \leq .001$.

Whiteman and Deutsch (1968), (*b*) selective school placement with children from poorer backgrounds going to "worse" schools and therefore learning less, or (*c*) unique achievement variance associated with social class and welfare status.

The final lines of Table 3 show the total $R^2\%$ and the multiple correlations as well as the $R^2\%$ and multiple correlation values corrected for shrinkage (McNemar, 1962), based on all predictors. The shrinkage is not great due to the relatively small number of predictors in relation to sample size. The three criteria are about equally well predicted. Proportions (corrected) of the total variance accounted for range from 41% to 43%.

DISCUSSION

The finding that performance on cognitive tasks contains two sources of variability, namely social-emotional and cognitive functioning, is in line with the position taken by Ziegler (1966) that performance on intelligence tests may be conceptualized as reflecting three distinct factors: (*a*) formal cognitive process, (*b*) informational achievement (content rather than the formal properties of cognition), and (*c*) personality variables.

A number of recent studies (Kohlberg, 1968; Levenstein, 1969; Massimo & Shore, 1963; Ziegler & Butterfield, 1968) have yielded evidence that gains in cognitive performance can be primarily due to the social-emotional components. What is needed is a clear delineation of those operations leading to changes in the cognitive components in contrast to those operations leading to changes in the noncognitive components of achievement measures.

The group of studies leading up to and including the present work suggest ways in which social-emotional functioning may affect cognition. It has been theorized (Kohn & Rosman, 1973a) that Factor I (Interest-Participation versus Apathy-Withdrawal) acts on cognitive functioning in the following ways:

1. The child who scores high on Interest-Participation learns more from his environment because of his curiosity, assertiveness,

and high rate of social interaction; the child high on Apathy-Withdrawal learns less from his environment because of his diminished contact with environmental stimuli and low rate of social interaction.

2. Children high on Interest-Participation are mentally more alert and more likely to engage in active thought processes such as hypothesis formation and testing, attention, discrimination, etc. Apathetic-Withdrawn children are mentally more inert and less inclined to make sense about what goes on around them; in fact, they may avoid thinking.

As for the processes intervening between Factor III (Task Orientation) and cognitive functioning, hypotheses analogous to those formulated for Factor I seem plausible as follows:

1. The child who scores high on Task Orientation learns more from his environment because he is organized and systematic in his contacts; the child scoring low on Task Orientation, because of his hyperactivity and restlessness, does not have sufficiently prolonged contact with specific objects and events in his environment to learn anything about their properties.

2. The child high on Task Orientation is more organized not only in the way he contacts his environment but also in the way he thinks; the child whose behavior is disorganized, hyperactive, and restless has thought processes that share these characteristics.

In summary, the child most likely to utilize his cognitive processes in a productive way and to make gains in cognitive achievement is one who is active, assertive, curious, task involved, and well organized. Further research is needed to show that this kind of child does achieve in school and that, in fact, children scoring high on Interest-Participation and Task Orientation do exhibit these behaviors in the classroom and do engage in these kinds of thought processes.

Components of Social-Emotional Functioning

Social-emotional functioning relevant to intellectual functioning is hypothesized to consist of two components: (*a*) a personality predisposition, that is, a person-specific component and (*b*) a situation-specific component.

In previous research, the factor dimensions showed moderate stability longitudinally (Kohn & Rosman, 1972a) and across settings (Kohn & Rosman, 1973b). The present study demonstrated significant correlations between preschool personality functioning and second-grade achievement—further evidence of the presence of a person-specific component. This component suggests that over a relatively wide range of learning conditions, children who are well adjusted on social-emotional Factors I or III or both learn more, and children who are disturbed on these dimensions learn less.

A previous study (Kohn, 1968) provided indirect evidence for a situation-specific component. The study showed large center-to-center differences in the average amount of Apathy-Withdrawal and Anger-Defiance, differences that could not be accounted for by differences in the background of the children in these centers.

Prescott and Jones' (1967) analysis of day care environments suggests some reasons for large center-to-center differences. Their data indicate that children's interest was a function of the encouragement, approval, and nurturance provided by the teachers as well as of the extent to which the lessons were designed to encourage creativity, experimentation, "pleasure, awe, and wonder." These findings imply that classroom and learning environments can be created to arouse optimal degrees of Interest-Participation.

The situational and person-specific components of social-emotional functioning suggest two types of strategies for intervention to improve cognitive performance, one directed at creating optimal learning environments and the second designed to improve the person-specific component of the child's social-emotional functioning.

Intervention aimed at the person-specific component would involve a therapeutic approach for the child who is disturbed in learning-relevant ways, that is, high on Ap-

athy-Withdrawal and/or low in Task Orientation.

A pilot study (Kohn & Rosman, 1971) has suggested that the children high in Apathy-Withdrawal have experienced a great deal of parental control and are made fearful in contacting the environment; children low in Task Orientation have experienced rejection, neglect, and a disorganized family life. This pilot study suggests that long sustained corrective emotional experiences would seem to be in order for both kinds of children in order to remediate both their personality disturbances and their cognitive functioning.

REFERENCES

COHEN, J. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 1968, **70**, 426–443.

KOHLBERG, L. Early education: A cognitive-developmental view. *Child Development*, 1968, **39**, 1013–1062.

KOHN, M. Competence and symptom factors in the preschool child. Unpublished manuscript, William Alanson White Institute (New York), 1968.

KOHN, M., & ROSMAN, B. L. Therapeutic intervention with disturbed children in day care: Implications of the deprivation hypothesis. *Child Care Quarterly*, 1971, **1**, 21–46.

KOHN, M., & ROSMAN, B. L. A social competence scale and symptom checklist for the preschool child: Factor dimensions, their cross-instrument generality and longitudinal persistence. *Developmental Psychology*, 1972, **6**, 430–444. (a)

KOHN, M., & ROSMAN, B. L. Relationship of preschool social-emotional functioning to later intellectual achievement. *Developmental Psychology*, 1972, **6**, 445–452. (b)

KOHN, M., & ROSMAN, B. L. Cognitive functioning in five-year-old boys as related to social-emotional and background-demographic variables. *Developmental Psychology*, 1973, **8**, 277–294. (a)

KOHN, M., & ROSMAN, B. L. Cross-situational and longitudinal stability of social-emotional functioning in young children. *Child Development*, 1973, **44**, 721–727. (b)

KOHN, M., & ROSMAN, B. L. A two-factor model of emotional disturbance in the young child: Validity and screening efficiency. *Journal of Child Psychology and Psychiatry*, 1973, **14**, 31–56. (c)

LEVENSTEIN, P. *Individual variation among preschoolers in a cognitive intervention program in low income families.* Mineola, N.Y.: Family Service Association of Nassau County, Inc., 1969. (ERIC Document Reproduction Service, ED 059 787.)

MASSIMO, J., & SHORE, M. The effectiveness of a comprehensive vocationally oriented psychotherapy for adolescent boys. *American Journal of Orthopsychiatry*, 1963, **36**, 634–642.

MCNEMAR, Q. *Psychological statistics.* New York: Wiley, 1962.

PEARSON, G. A survey of learning difficulties in children. *The Psychoanalytic Study of the Child*, 1952, **7**, 322–386.

PEARSON, G. *Psychoanalysis and the education of the child.* New York: Norton, 1954.

PRESCOTT, E., & JONES, E. Group day care as a child-rearing environment. Unpublished manuscript, Pacific Oaks College, 1967.

STOTT, L. H., & BALL, R. S. Infant and preschool mental tests: Review and Evaluation. *Monographs of the Society for Research in Child Development*, 1965, **30**(3).

WECHSLER, D. Intelligence: Definition, theory and the IQ. In R. Cancro (Ed.), *Intelligence—Genetic and environmental influences.* New York: Grune & Stratton, 1971.

WHITEMAN, M., & DEUTSCH, M. Social disadvantage as related to intellective and language development. In M. Deutsch, I. Katz, & A. R. Jensen (Eds.), *Social class, race, and psychological development.* New York: Holt, Rinehart & Winston, 1968.

ZIEGLER, E. Mental retardation: Current issues and approaches. In L. W. Hoffman & M. L. Hoffman (Eds.), *Review of child development research.* Vol. 2. New York: Russell Sage Foundation, 1966.

ZIEGLER, E., & BUTTERFIELD, E. C. Motivational aspects of changes in IQ test performance of culturally deprived nursery school children. *Child Development*, 1968, **39**, 1–15.

(Received March 26, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 2, 277–284

DETERMINATION OF THE ORDERING AMONG SEVEN PIAGETIAN TASKS BY AN ORDERING-THEORETIC METHOD[1]

WILLIAM M. BART[2]
University of Minnesota

PETER W. AIRASIAN
Boston College

This study applied a method of generating task hierarchies (ordering theory) to seven Piagetian tasks. Thirty subjects were individually administered three concrete operational and four formal operational tasks. Analysis of the response patterns on the dichotomously scored tasks revealed the following findings. First, the Piagetian theory that success on concrete operational tasks is a necessary prerequisite to success on formal operational tasks was confirmed. Second, the seven tasks were closely interrelated with an array of prerequisite relations more complex than a simple linear hierarchy. The methodology utilized appears to have value for defining nonlinear lines of implication among behavioral science phenomena.

The purpose of this study was to explicate ordering theory, a measurement model which identifies necessary and sufficient performance conditions between test items or tasks, by analyzing seven Piagetian tasks. Ordering theory is an extension of Guttman (1944, 1950) scaling procedures in that it identifies nonlinear as well as linear hierarchies among items or tasks. Scalogram analysis is used to order a group of items or tasks into a linear hierarchy and to evaluate whether or not the hierarchy is unidimensional and cumulatively hierarchical. The degree to which a group of items or tasks is judged to be unidimensional and cumulative is determined by the extent to which "passes" (scores of 1) on any item or task co-occur with passes on all items or tasks ranked as less difficult. The inverse is also true; that is, a hierarchy is unidimensional and cumulative insofar as "failures" (scores of 0) on an item or task co-occur with failures on items ranked as more difficult. Scalogram analysis, however, is constrained to the identification of linear hierarchies. A linear hierarchy is one in which a series of test items or tasks are arranged in a hierarchy such that one and only one item or task appears at any given level of the hierarchy and one and only one task is immediately prerequisite to or a sufficient condition for any other item or task.

Ordering theory is a measurement model which is based upon scalogram analysis but which extends scalogram techniques to nonlinear item or task hierarchies. Ordering theory is an approach to fundamental measurement which has as its primary intent either the determination of a hierarchy for a set of test items or tasks or the testing of a hypothesized hierarchy among a set of items or tasks (Airasian & Bart, 1973). At a foundational level of explanation, ordering theory is a measurement model with a Boolean algebraic framework in which item or task response patterns are viewed as at-

[1] This work was supported in part by grants to the University of Minnesota Center for Research in Human Learning from the National Science Foundation (GS-1761), the National Institute of Child Health and Human Development (HD-01136-04), and from the Graduate School of the University of Minnesota. The research reported herein was also supported by a grant (OE-09-332189-4533-32) from the Bureau of Education for the Handicapped, U. S. Office of Education, to the Research and Development Center in Education of Handicapped Children.

[2] Requests for reprints should be sent to William M. Bart, Psychological Foundations of Education, 330 Burton Hall, University of Minnesota, Minneapolis, Minnesota 55455.

oms in a Boolean algebra with as many generators as there are items or tasks being considered (Goodstein, 1963). Though there are a variety of ways of articulating ordering theory, it is probably most meaningfully classified as a deterministic measurement model which is useful for identifying both linear and nonlinear, qualitative prerequisite relations among test items or tasks. In this regard, it is a more general case of scalogram analysis.

Ordering-theoretic procedures were used in this study to determine prerequisite relations between pairs of Piagetian tasks. The prerequisite relation is considered here, since that type of relation is a primary interest to behavioral scientists, especially in their quest to identify causal relationships among phenomena.

An item i is a prerequisite to an item j to the extent that the (0,1) response pattern, where 0 represents the score on item i and 1 represents the score on item j, occurs infrequently. The (0,1) response pattern is viewed as a disconfirmation that a correct response to item i is a prerequisite to a correct response to item j. In defining prerequisite relations between items, ordering theory shares one limitation with scalogram analysis. Both measurement models are deterministic rather than probabilistic. Thus, neither model incorporates a method of dealing with the probability of encountering random error in item response patterns. As a consequence of this limitation, ordering-theoretic analyses rely upon the use of a preset tolerance level for error. The tolerance level sets the number of disconfirmatory response patterns which will be accepted in defining a prerequisite relation between two items. Thus, for a 5% tolerance level and n subjects, one would tolerate at most .05 n disconfirmatory response patterns between items in an item pair before accepting a prerequisite relation.

The logical relation between any two items such as the prerequisite relation can thus be identified from an examination of the magnitude of the cell frequencies in the 2×2 matrix relating the item scores. Thus, for a 5% tolerance level and n subjects reacting to items i and j, an absence of subjects (i.e., $\leqslant .05n$) in only the i j 01 cell indicates that i is a logical precondition of j, an absence of subjects (i.e., $\leqslant .05n$) in the i j 01 and 10 cells indicates that items i and j are logically equivalent, and the presence of subjects (i.e., $>.05n$) in all cells indicates that items i and j are logically independent. The logical relation of prerequisite and its related logical relations of equivalence and independence are the crucial relations used in ordering theory.

There are two general strategies for the implementation of tolerance levels in examining the hierarchical structure or ordering among a set of items. Within one strategy, a hierarchy with its array of prerequisite relations is hypothesized for a set of items or tasks. The response patterns that would be disconfirmatory for the entire hierarchy are identified and a tolerance level is established (Airasian, 1971a, 1971b). The hypothesized hierarchy is then accepted if the frequency of obtained disconfirmatory response patterns is less than or equal to the prescribed tolerance level.

An alternative strategy, used when no a priori hierarchy among items is hypothesized, identifies prerequisite relationships between item pairs. In this strategy, all possible item pairs are investigated to identify prerequisite relations. The prerequisite relation for a particular pair of items is accepted if the frequency of obtained disconfirmatory response patterns for the item pair is less than or equal to the frequency of such response patterns established by the tolerance level. This procedure is followed to test each of the possible hypothesized prerequisite relations with the same tolerance level being used for each testing. Within this strategy, item response patterns that disconfirm one prerequisite relation may be different from the item response patterns that disconfirm another prerequisite relation.

Several discussions of ordering theory have been provided. Airasian and Bart (1973) articulated the general nature of ordering theory. Bart and Krus (1973) described an ordering-theoretic technique by which item or task hierarchies could be determined. Airasian, Bart and Greaney

(1973) investigated a propositional logic game and compared the ordering-theoretic results to the results of scalogram analyses. Two defining properties or ordering theory cited in these discussions are that all test items or tasks to be examined must be dichotomously scored and that all subjects in a sample must respond to all of the items or tasks.

A computer program which searches out all of the prerequisite relations between all item pairs has been developed (Lele & Bart, 1971). The program accepts a prerequisite relation between an item pair if the frequency of obtained disconfirmatory response patterns is less than or equal to the frequency of such response patterns established by the tolerance. This procedure is followed to test for a prerequisite relation between all possible item pairs, with the same tolerance level used for each testing. In this study, the computer program was used to examine the hierarchy among seven Piagetian tasks that tested for various levels of operational reasoning.

Method

Subjects

The subjects were 30 high school freshmen, 16 males and 14 females, whose mean age was 14.1 years, with a range from 13.0 years to 15.1 years. The subjects were enrolled in a high school in the Chicago area.

Procedure

Seven tasks that tested for operational schemes which are general internal forms of knowing (Furth, 1967) were administered to each of the 30 subjects in the following manner: each subject was tested individually with the seven tasks being given one at a time in a prescribed order over a 75-minute period. The order of task presentation was as follows:

1. matrix task,
2. tactile seriation task,
3. animal classification task,
4. equilibrium in the balance task,
5. projection of shadows task,
6. oscillation of the pendulum task, and
7. conservation of motion in a horizontal plane task.

The first three tasks tested concrete reasoning and were derived from *The Early Growth of Logic in the Child* (Inhelder & Piaget, 1969). The other four tasks tested formal reasoning and were derived from *The Growth of Logical Thinking from Childhood to Adolescence* (Inhelder & Piaget, 1958). For each task, the responses of each subject were independently judged and scored by two raters.

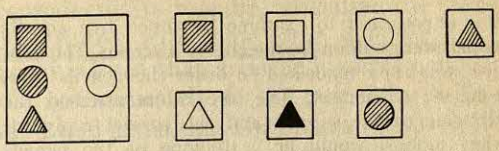


FIGURE 1. Eight cards used in the matrix task.

1. *Matrix task.* The matrix task involved the use of eight cards. On one card, there were five figures and on each of the other cards there was only one figure. The eight cards with their figures are shown in Figure 1. This display is identical to the way in which the cards were shown to each subject.

In this task, the subject was shown the eight cards with their circle, triangle, and square figures and then was asked

> Which of the seven figures here [the experimenter points to the figures on the seven small cards] best fits in this blank space [the experimenter points to the blank space on the large card] so that it fits this way (horizontally) and that way (vertically)?

The correct answer is the figure on the small card in the second row and the first column of small card array. The subject was then asked "Why did you choose that one?" The correct answer for this question would indicate that the correct figure is a triangle because the row that it is in has a triangle, that the figure is uncolored because the column that it is in has only uncolored figures, and that the rows are determined by shape and the columns are determined by color or degree of shading. After the subject's response to that question was assessed, the subject was then asked "Tell me whether any of the other figures might fit in as well or better?" This question tested the stability of the choice of figure indicated in the response to the first question. The correct answer to this question was basically that no other figure fitted in the space well.

The two defined scores for this task were 1 and 0, with 1 being given if the sum of the questions answered correctly according to the two raters was either 5 or 6 and 0 being given if that sum was less than 5. A score of 1 for this task was an indication that a subject had the scheme of multiplicative classification which is proper to the period of concrete operations. For this task, the raters had point value agreement of 86 of the 90 question responses. The percentage of interrater agreement for the question responses for this task was used as an index of interrater reliability and had the value of .96.

2. *Tactile seriation task.* In this task, the subject

was encouraged to examine by touch five pencils which were hidden from sight by a screen. The subject was then requested to order the pencils from longest to shortest. The two raters assessed the drawing of the subject and the correct reply for that request would be a drawing of the pencils such that the bases of the pencil figures would be on the same horizontal line and the pencils would be in order from longest to shortest with 5 pencil figures being indicated.

The defined scores for this task were 1 and 0, with 1 being given if the subject correctly replied to both task assignments according to both raters and 0 being given otherwise. A score of 1 for this task was an indication that a subject had the scheme of simple seriation which is proper to the period of concrete operations. For this task, the interrater agreement was 1.00.

3. *Animal classification task.* In this task, the subject was shown 12 cards with a different animal picture on each card. Four ducks, three other birds (chicken, sparrow, and parrot), and five animals that are not birds (mouse, fish, horse, poodle, and snake) were depicted on the cards. The subject was then asked to make separate piles, placing the animals which were like each other in a pile. After that activity, the subject was asked the following questions in this order:

> Are there more ducks or more birds?
> Are there more birds or more animals?
> If a fox ate all the birds, would there be any ducks left?
> What if one killed all the animals, would there be any birds left?

Two raters independently assessed the responses to these questions as to their correctness; the correct answers were "more birds," "More animals," "no," and "no" to the four questions, respectively. The two defined scores for this task were 1 and 0, with 1 being given if the sum of questions answered correctly according to the two raters was either 7 or 8 and 0 being given if that sum was less than 7. A score of 1 on this task indicated that a subject had the scheme of additive classification which is proper to the period of concrete operations. For this task, perfect interrater agreement was reached.

4. *Equilibrium in the balance task.* In this task, a conventional balance with an array of six weights was used. Inhelder and Piaget's (1958) procedures for questioning the subject and for assessing the responses for this task were used in this study. Thus, subjects were assessed to be at various substages: 1-A, 1-B, 2-A, 2-B, 3-A, and 3-B, with Substage 3-B being the highest substage as a consolidated level in the period of formal operations. The raters independently awarded points to subject responses according to the following convention: 1 point for substage 1-A and 1 additional point for each successive substage with 6 points for substage 3-B. The two defined scores for this task were 1 and 0, with 1 being given if the subject received a combined rater point total of less than 10. Since the array of substages used in the other three formal reasoning tasks were very similar in form to the substages for this task, the same quantification procedure for the attribution of task scores was used for the other three tasks. A score of 1 on this task indicated that a subject had the scheme of proportionality which is proper to the period of formal operations. For this task, the rater agreement was .87.

5. *The projections of shadows task.* This task involved a screen, an electric light, a baseboard with a sequence of holes set at one-inch distances, and four rings with diameters of one, two, three, and four inches, respectively. Basically, the task of the subject was to indicate what distance from the light a given ring should be placed so that its shadow on the screen conformed to the shadow produced by another ring placed at a given distance. Inhelder and Piaget's (1958) procedures for interviewing the subject and for assessing the responses for this task were used in this study. The score of 1 for this task indicated that a subject had the scheme of proportionality which is proper to the period of formal operations. For this task, interrater agreement was .90.

6. *The oscillations of the pendulum task.* This task consisted of a string which could be lengthened or shortened, a wooden apparatus to hang the string, and a set of weights. Basically, the task of the subject was to identify the factor that regulated the rate of oscillation of the pendulum. Inhelder and Piaget's (1958) procedures for questioning the subject and for assessing the responses for this task were used. A score of 1 for this task indicated that a subject had the scheme of combinatorial reasoning which is proper to the period of formal reasoning. The interrater agreement for this task was .87.

7. *Conservation of motion in a horizontal plane task.* This task involved a long board with a central groove, a thick plastic pendulum hammer fixed at one end of the board, and six balls, each of which would be projected down the groove by the action of the hammer head. Basically, the task of the subject was to identify the factors that regulated the distances balls traveled down the groove when struck with a constant force by the hammer. Inhelder and Piaget's (1958) procedures for questioning the subject and for assessing the responses for this task were used in this study. A score of 1 for this task indicated that a subject had the scheme of combinatorial reasoning which is proper to the period of formal operations. For this task, interrater agreement was .83. Though this value was the lowest measure of interrater reliability, it still was quite high and thus substantial interrater reliability was established for this task.

For each of the seven tasks used, procedures for the dichotomous scoring of the task responses were thus established. From the employment of these tasks and the highly reliable response assessment procedure, a 30 × 7 task score matrix with 1s and 0s was generated with the rows

TABLE 1

ITEM RESPONSE PATTERNS FOR SEVEN PIAGETIAN TASKS FOR A SAMPLE OF 30 SUBJECTS

Item response patterns	Task							Frequency
	1	2	3	4	5	6	7	
1	0	0	0	0	0	0	0	1
2	0	0	1	0	0	0	0	1
3	1	1	1	0	0	0	0	3
4	1	1	1	1	0	0	0	7
5	1	1	1	0	0	1	0	7
6	1	1	1	1	0	1	0	6
7	1	1	1	1	1	1	1	5

referring to the 30 subjects and the columns referring to the seven tasks.

RESULTS

The task score matrix for the 30 subjects was generated and was used in an array of analyses that provided some comparisons of ordering-theoretic techniques with other more conventional data analytic techniques. One basic use of the task score matrix was to identify the basic item response patterns that occurred.

As can be seen in Table 1, seven distinct item response patterns were generated by the 30 subjects. In addition, the subjects tended to perform fairly well on the tasks since most of the total task scores for the subjects were more than 3.

A scalogram analysis was performed on the item response patterns shown in Table 1. For the data in Table 1, the coefficient of reproducibility was .93, indicating the tasks constituted, with some error, a linear hierarchy. Given that a linear hierarchy model did not accurately fit the task data in Table 1 and given that Guttman's (1950) scalogram analysis is primarily limited to the fitting of item data to a linear hierarchical model, ordering-theoretic procedures were used to determine the ordering among the Piagetian tasks.

From Table 1, the frequencies in the 01 and 10 cells for each pair of tasks were determined and reported in Table 2. As has been noted, those frequencies are crucial in determining the logical relations among the tasks. Those logical relations between tasks are also indicated in Table 2. Thus, for example, one can infer from Table 2 that Tasks 4 and 6 are logically equivalent.

If tolerance levels are established, then the tolerance levels will determine the maximal frequencies either in the 01 cell or in the 10 cell that may occur in order for the prerequisite relation to be accepted for any given task pair. Though the ordering-theoretic program used to analyze this task data does not actually generate the inter-task score contingency tables needed for the ordering-theoretic data analysis to occur, those contingency tables could be used to formulate the hierarchy among the tasks for any tolerance level in accordance with the rules for determining prerequisite relations that have been discussed.

With the use of a computer program, the

TABLE 2

NUMBER OF SUBJECTS IN THE 10 AND 01 CELLS BETWEEN EACH OF THE SEVEN OPERATIVITY TASKS (THE NUMBER OF SUBJECTS WHO FOUND ONE TASK MORE DIFFICULT THAN THE OTHER TASK)

Task	Task[a]											
	2		3		4		5		6		7	
	10	01	10	01	10	01	10	01	10	01	10	01
1. Matrix	0	0[b]	0	1	10	0	23	0	10	0	23	0
2. Seriation			0	1	10	0	23	0	10	0	23	0
3. Animal					11	0	24	0	11	0	24	0
4. Balance							13	0	7	7[c]	13	0
5. Shadow									0	13	0	0[b]
6. Pendulum											13	0

[a] Zero in only the 01 or only 10 cell, but not both, indicates precondition relationship between tasks.
[b] Pattern indicates logical equivalence between tasks.
[c] Pattern indicates logical independence of tasks.

task response patterns were subjected to an ordering-theoretic analysis with a 1% tolerance level being established. Hence, pairs of tasks, where greater than 1% disconfirmatory response patterns were found, were rejected as manifesting a prerequisite relation. In the case of 30 subjects, a 1% tolerance level established a frequency of .03 as the maximal frequency that either the 01 cell or the 10 cell can manifest before the prerequisite relation is rejected. Upon examination of disconfirmatory frequencies set by various tolerance levels for 30 subjects, a 1% tolerance level and a 0% tolerance level were equivalent for 30 subjects in that no disconfirmatory response patterns were allowed before corresponding prerequisite relations were rejected. As a consequence, the identified ordering manifested perfect reproducibility. In essence, then, the response patterns for each student on the seven Piagetian tasks were examined to determine which tasks were prerequisites for the satisfactory accomplishment of other tasks. Figure 2 depicts the ordering determined for the seven tasks. Note that the ordering is similar to a hierarchy.

Figure 2 indicates prerequisite relations as well as relations of logical equivalence and logical independence between tasks in various task pairs. Success on one task is a prerequisite to success on another task if the number representing the first task is connected to the number representing the second task by a single arrow →. For the relation, "is a prerequisite to," phrases such as "is a precondition for" and "is a necessary condition for" may be viewed as being synonymous. Also, if success on Task A is a necessary condition for success on Task B, then success on Task B is a sufficient condition for success on Task A. In other words, the converse of the relation, "is a necessary condition for," is the relation, "is a sufficient condition for." When both of those relations hold between two tasks, then the relation of "is a necessary and sufficient condition for," which is its own converse, holds between the two tasks. These points are made merely to indicate that there are many ways of articulating the prerequisite relation between tasks. The behavioral researcher may use the way of articulating a specific logical relation between two tasks that would be most meaningful to him. An example of a prerequisite relation among the seven tasks as indicated in Figure 2 is that success on Task 1 is a prerequisite to success on Task 4.

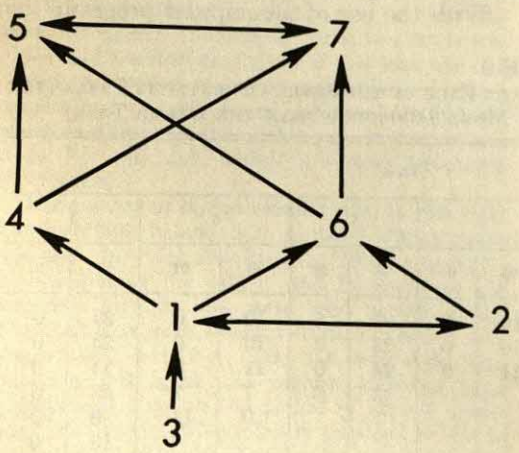


FIGURE 2. Ordering diagram for seven piagetian tasks. (Double arrow ↔ indicates logical equivalence and single arrow → indicates logical precondition.)

A task is logically equivalent to another task if success on one task co-occurs only with success on the other task. In other words, the task is logically equivalent to another task if success on the first task is a necessary and sufficient condition for success on the other task. Logical equivalence between two tasks indicates that the task score for one task is the same as the task score for the other task for any tested subject. A figurative representation for logical equivalence between two tasks entails a double arrow ↔ connecting numbers which are referents for the two tasks. An example of two logically equivalent tasks as indicated in Figure 2 are Tasks 1 and 2.

A task is logically independent of another task if the score for one task is unrelated to the score for the other task. Logical independence between two tasks indicated that each of the four possible response patterns for the two tasks occurs at a frequency greater than that established by a tolerance level. Figuratively, one task is logically independent of another task if the number representing the first task is not connected

to the number representing the second task by a line that passes either in a general upward direction or in a general downward direction or horizontally. An example of two logically independent tasks as indicated in Figure 2 are Tasks 4 and 6. With the logical relations between tasks identified, the item ordering can be constructed as indicated in the case of Figure 2. However, the ordering determined for a set of tasks will tend to vary with the tolerance level employed. For example, if a 5% tolerance level were used, the Figure 2 structure would be different with the graphical relation

$$1 \leftrightarrow 2 \leftrightarrow 3$$

then occurring for Tasks 1, 2, and 3. A 1% tolerance level was used primarily on the grounds that the experimenters wanted to be as strict as possible as they allowed for a minimal amount of error. Also, in identifying prerequisite relations, the frequencies between 01 and 10 cells will differ and in most cases the difference will be significant. However, if one employs a low tolerance level such as 1%, a small difference between the 01 and 10 cells frequencies for two tasks may occur with a logical precondition relation holding for the two tasks; this situation occurred for Tasks 1 and 3 and for Tasks 1 and 2. In general, with higher tolerance levels, a precondition relation implies that it is more likely for one task to be significantly more difficult than the other task. With lower tolerance levels, logical equivalence between two tasks implies that it is more likely for the difficulties of the two tasks to be equal.

One prominent finding from Figure 2 and Table 2 is that evidence is provided that the concrete operational schemes required by Tasks 1–3 are necessary but not sufficient conditions for the formal operational schemes required by Tasks 4–7. In other words, for the subjects tested, classification and seriation schemes were prerequisites for the formal operational schemes of proportionality and combinatorial reasoning. That finding is compatible with Piagetian theory insofar as Piaget (e.g., Inhelder & Piaget, 1958) in a host of books has contended that concrete reasoning is a precondition to formal reasoning.

Another prominent finding is that evidence is provided that the seven tasks are closely interrelated with an array of prerequisite relations indicating a hierarchical structure more complex than that of a simple linear hierarchy. If the seven tasks were in a simple linear hierarchy, then any two tasks would be related by the prerequisite relation and there would be no two tasks that are logically independent. However, Tasks 4 and 6 were found to be logically independent for the sample tested. Within the hierarchy of the seven tasks, the two tasks (4 and 5) that assess the scheme of proportionality, the two tasks (6 and 7) that assess the scheme of combinatorial reasoning, and the two tasks (1 and 3) that assess classification skills each determine a linear scale.

Discussion

The purpose of this research was not to determine the universal and eternal ordering of the seven Piagetian tasks but rather to demonstrate how ordering-theoretic procedures can be used to determine the pattern of logical relations among a class of Piagetian cognitive processes—in this case, the schemes required by seven Piagetian tasks. The ordering determined for the seven Piagetian tasks used held for the 30 subjects tested and was compatible with Piagetian theory, but many more samples of subjects from around the world would have to be assessed with the seven tasks and the data analyzed with ordering-theoretic procedures in order to establish the level of universality and population generalizability that the ordering has.

However, it should be clear that the possible use of ordering-theoretic methods for education and other social sciences are myriad. Whenever logical relations between phenomena or events are of interest to a researcher, ordering theory can and should be used. Traditional psychometric theory has given us a family of techniques that allow us to discern patterns of association in terms of correlations. However, such techniques do not greatly help us to deter-

mine lines of causation. Ordering theory can reveal lines of implications among phenomena, which in turn can serve as a basis for hypothesizing lines of causation to be tested in experimental settings. Ordering theory should be a useful tool to the educator or social scientist.

An example of the educational utility of ordering theory lies in its potential relationship to the task analysis method of Gagne (1965). Basically, some primary task competency is designated to be an important educational objective and then a hierarchy of prerequisite tasks underlying the primary task is generated in an a priori manner. One intent of such a task hierarchy is that it indicates a path by which educational development can proceed with the end being the attainment of the primary task competency. Thus, the Gagne method has utility for formulating instructional and curriculum-sequencing decisions. The functions of ordering theory for the Gagne method of task analysis lie in the empirical determination of task hierarchies which can be guides for instructional and curricular decisions and in the empirical testing of a hypothesized task hierarchy (Airasian, 1971a). Thus, ordering theory would be an aid to the fields of curriculum and instruction.

As for developmental psychology and Piagetian theory and research, ordering theory should be helpful in providing information on hierarchical structures among behaviors which, in turn, constitute a common and useful way to conceptualize behaviors from a developmental viewpoint. In the case of this study, the hierarchy among the cognitive behaviors assessed was very compatible with the hierarchy among the same cognitive behaviors which have been posited by Piaget. It is anticipated that other cognitive behavior hierarchies posited by Piaget will be empirically verified using ordering-theoretic methods.

REFERENCES

Airasian, P. W. A method for validating sequential instructional hierarchies. *Educational Technology*, 1971, **11**, 54–56. (a)

Airasian, P. W. A study of behaviorally dependent, classroom taught task hierarchies. (Educational Technology Research Report Series 3) Englewood Cliffs, N. J.: Educational Technology Publications, 1971. (b)

Airasian, P. W., & Bart, W. M. Ordering theory: A new and useful measurement model. *Educational Technology*, 1973, **13**, 56–60.

Airasian, P. W., Bart, W. M., & Greaney, B. J. An ordering-theoretic analysis of a propositional logic game. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

Bart, W. M., & Krus, D. J. An ordering-theoretic method to determine hierarchies among items. *Educational and Psychological Measurement*, 1973, **33**, 291–300.

Furth, H. G. *Piaget and knowledge: Theoretical foundations*. Englewood Cliffs, N. J.: Prentice-Hall, 1967.

Gagne, R. M. *The conditions of learning*. New York: Holt, Rinehart & Winston, 1965.

Goodstein, R. L. *Boolean algebra*. London: Pergamon Press, 1963.

Guttman, L. A. A basis for scaling qualitative data. *American Sociological Review*, 1944, **9**, 139–150.

Guttman, L. A. A basis for scalogram analysis. In S. A. Stouffer et al., *Studies in social psychology in World War II*. Vol. 4. *Measurement and prediction*. Princeton: Princeton University Press, 1950.

Inhelder, B., & Piaget, J. *The growth of logical thinking from childhood to adolescence*. New York: Basic Books, 1958.

Inhelder, B., & Piaget, J. *The early growth of logic in the child*. New York: Norton, 1969.

Lele, K., & Bart, W. M. *Preliminary item analysis by ordering theory*. Minneapolis: University of Minnesota Research and Development Center in Education of Handicapped Children, 1971. (Mimeo)

(Received April 3, 1973)

CHANDLER

CHANGING CLASSROOM BEHAVIOR, 2nd Ed.

Merle L. Meacham, University of Washington and Allen E. Wiesen

This new edition of a widely used text is based on the premise that an educated individual is one who is self-directed and equipped with skills that give him a measure of control over his environment. To achieve this goal, the authors advocate "objective teaching," a method that directly applies such approaches as reinforcement, extinction, imitative learning, and programming. Many practical illustrations show him to use these and related techniques in class.

February, 1974 264 pages paper, $4.95

THE ADOLESCENT: Case Studies for Analysis

Nancy C. Ralston and G. Patience Thomas, both of the University of Cincinnati

Ten case studies provide insight into the varied problems of adolescence. Introductory chapters explain the case study method —its philosophy, uses, and limitations—and summarize basic philosophical theories of human learning, development, and behavior. Each case is followed by questions for analysis and discussion.

March, 1974 208 pages paper, $4.95

THE CHILD: Case Studies for Analysis

Nancy C. Ralston and G. Patience Thomas, both of the University of Cincinnati

Case studies with interpretative questions that will help students develop a broad understanding of some of the typical behavior patterns of childhood, along with insight into the causes and meaning of different kinds of behavior. A sample case study outline is included.

1972 200 pages paper, $3.95

Just published in paperback—

THE DIAGNOSIS OF REASONING IN THE MENTALLY RETARDED, 2nd Ed.

Bärbel Inhelder
Preface by Jean Piaget
Translated by Will Beth Stephens

The first paperback edition of a classic work. Inhelder's book remains the major research effort to relate Piaget's theory of the development of intelligence to the area of mental retardation. "Findings from this study will continue to serve for years to come as a framework for research in the area of retardation."

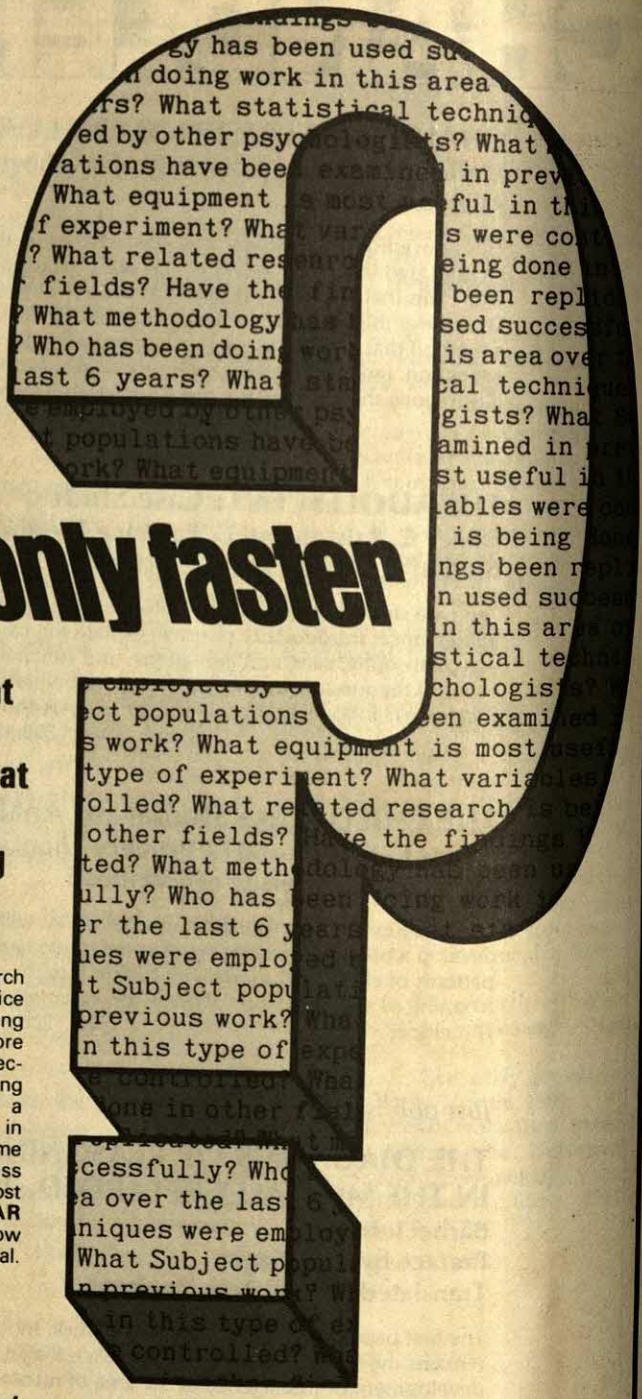
—Library Journal

1968 368 pages paper, $7.95

An Intext Publisher

CHANDLER PUBLISHING COMPANY • 257 PARK AVENUE SOUTH • NEW YORK, N.Y. 10010

If you can always find the psychological information you're looking for in the library stacks . . .



Pasar is only faster

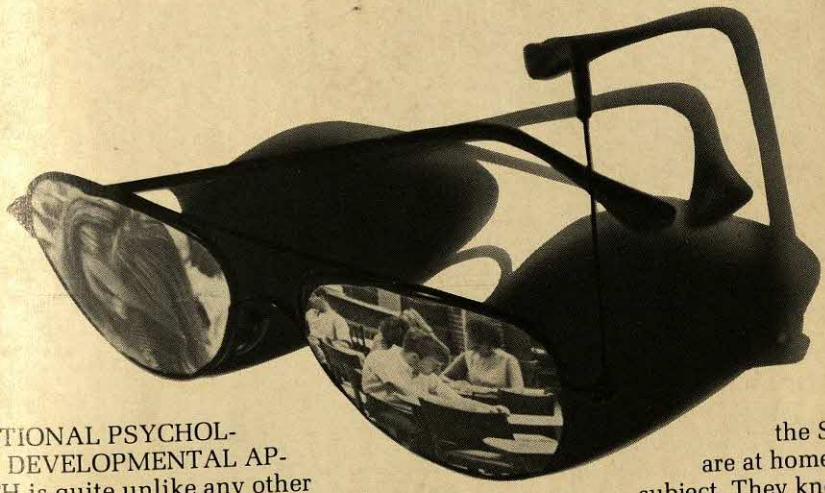
but chances are great that Pasar will enable you to retrieve some material that would remain lost to the manual searcher devoting months to the same task.

PASAR (*Psychological Abstracts* Search and Retrieval) is an automated service that affords rapid bibliographic searching at a level of complexity never before possible. A computer printout of *PA* records (from 1967 to the present) satisfying the information need you specify on a **PASAR** Request Form is sent to you in about 2-3 weeks—a fraction of the time frequently required to perform a less satisfactory manual search. Average cost ranges $40-$60. To obtain a **PASAR** Request Form, write to the address below or consult a recent issue of the *PA* journal.

Psychological Abstracts Information Services
AMERICAN PSYCHOLOGICAL ASSOCIATION
1200 Seventeenth Street, N.W. Washington, D.C. 20036

You don't actually read Sprinthall & Sprinthall.

You see with it.



EDUCATIONAL PSYCHOLOGY: A DEVELOPMENTAL APPROACH is quite unlike any other text you've ever seen. In a lot of ways, it isn't a textbook at all. But then, Richard Sprinthall and Norman Sprinthall set out to create a different kind of course.

And they did.

They came up with a course that shows what educational psychology is all about. A course that lets the student experience ed psych.

The book is that good. That complete. That coherent. And viable. Loaded with special features.

But the one thing that puts it all in sharp focus is the Sprinthalls' writing style. It's lucid. Comfortable. Even entertaining.

Because the Sprinthalls are at home with their subject. They know how interesting it is. How enigmatic. And how important.

But they haven't become so preoccupied that they lose sight of the student. Consequently, the student won't lose sight of educational psychology.

As one reviewer put it, "The writing is quite clear and communicative, not pedantic or labored." And from another: "The writing provides fast, fluent, interesting reading; consistently appropriate for a first course."

Superb writing is just one aspect of the Sprinthall achievement. But it's the one that makes all the others clearly visible.



College Division
ADDISON-WESLEY PUBLISHING COMPANY, INC.
Reading, Massachusetts 01867

CHARLES C THOMAS • PUBLISHER

ADOLESCENT IMMATURITY: Prevention and Treatment *by* **I. Newton Kugelmass,** *Consultant to the Departments of Health and Hospitals, New York City.* This work unravels the broadening developmental complexities of adolescent maturation and the accompanying interlocking immaturities that characterize unstable youth. It provides a clinical spectrum of maturation processes for initiating the syntheses of sound practical approaches to the final milestone in the attainment of full-fledged adulthood. *'73, 320 pp., 2 il., 11 tables, $12.50*

THE MANAGEMENT OF CHILDHOOD BEHAVIOR PROBLEMS IN SCHOOL AND AT HOME. *Edited by* **Lloyd Keith Daniels,** *Central Connecticut State College, New Britain. (108 Contributors)* Designed for parents and educators, this text describes the principles and application of behavior modification for the management of children and young adults in the classroom and at home. The major issues the book is concerned with are the modification of avoidance and approach problems by the classroom teacher and the role of parents in the remediation process. *'74, about 476 pp. (7 x 10), 75 il., 22 tables*

CHILDREN'S EXPERIENCE WITH DEATH *by* **Rose Zeligs.** An authoritative book by a practicing psychologist for professional and lay people, parents and students who want to learn how the existence of death influences a child's life and how they can help the child deal with death in a constructive way. Chapters include developmental concepts of death; fear of death; the dying child himself and his hospital experiences; his reactions to being handicapped; his response to the loss of a parent; and the role of suicide in his life. *'74, about 276 pp.*

THE INSTITUTIONALIZED SEVERELY RETARDED: A Study of Activity and Interaction *by* **Richard H. Wills,** *Univ. of California, Berkeley.* Two important breakthroughs in mental retardation research are provided – a new classification of the severely retarded based on clear-cut and observable qualitative differences in behavioral ability and a long-needed comprehensive examination of severely retarded behavior with a precise detailed study of each type of activity and interaction that occurs among severely retarded residents on an institutional ward. *'73, 208 pp., 1 il., 16 tables, $9.75*

THE REHABILITATION MEDICINE SERVICES: An Orientation *by* **Laurence P. Ince,** *Center for Learning Disabilities, Briarwood, New York.* Written for the student planning to enter the field of rehabilitation or a health care profession, this book will acquaint him with the roles and functions of the treatment disciplines and how the disciplines interact with one another in their therapy functions. Discussed in detail are the disorders and disabilities of rehabilitation patients; physical therapy; occupational therapy; speech and language therapy and auditory training; psychology; social work; vocational counseling; recreation therapy; and nursing. *'74, about 672 pp. (6 3/4 x 9 3/4), 101 il.*

Orders with remittance sent, on approval, postpaid

301-327 East Lawrence Avenue • Springfield • Illinois • 62717

Journal of Educational Psychology

VOLUME 66 NUMBER 3

JUNE 1974

Ake
Docket Disc
to the Library
R. Dan
2/9/79

AMERICAN PSYCHOLOGICAL ASSOCIATION

EDITOR

JOANNA WILLIAMS, *Teachers College, Columbia University*

ASSOCIATE EDITOR

SAMUEL BALL, *Educational Testing Service*

ADVISORY EDITORS

PAUL R. AMMON, *University of California, Berkeley*
JERE BROPHY, *University of Texas at Austin*
C. VICTOR BUNDERSON, *Brigham Young University*
LEONARD S. CAHEN, *Educational Testing Service*
LOIS-ELLIN DATTA, *National Institute of Education*
FREDERICK B. DAVIS, *University of Pennsylvania*
FRANCIS J. DIVESTA, *Pennsylvania State University*
ROBERT EBEL, *Michigan State University*
DAVID ELKIND, *University of Rochester*
LAWRENCE T. FRASE, *Bell Laboratories*
N. L. GAGE, *Stanford University*
ROBERT GAGNÉ, *Florida State University*
ROBERT GLASER, *University of Pittsburgh*
MIRIAM L. GOLDBERG, *Teachers College, Columbia University*
EDMUND W. GORDON, *Teachers College, Columbia University*
VERNON C. HALL, *Syracuse University*
ARTHUR R. JENSEN, *University of California, Berkeley*
PHYLLIS A. KATZ, *City University of New York, Graduate School and University Center*
HERBERT J. KLAUSMEIER, *University of Wisconsin—Madison*
ROBERT E. LANA, *Temple University*
GERALD S. LESSER, *Harvard University*
JOEL R. LEVIN, *University of Wisconsin—Madison*
FREDERICK J. MCDONALD, *Educational Testing Service*
WILBERT J. MCKEACHIE, *University of Michigan*
FRANK B. MURRAY, *University of Delaware*
LAUREN B. RESNICK, *University of Pittsburgh*
WILLIAM D. ROHWER, JR., *University of California, Berkeley*
ROBERT ROSENTHAL, *Harvard University*
ERNST Z. ROTHKOPF, *Bell Laboratories*
S. JAY SAMUELS, *University of Minnesota*
SEYMOUR B. SARASON, *Yale University*
MELVYN I. SEMMEL, *Indiana University*
IRVING SIGEL, *Educational Testing Service*
HARRY SILBERMAN, *University of California, Los Angeles*
RICHARD E. SNOW, *Stanford University*
GEORGE G. THOMPSON, *Ohio State University*

JOURNAL STAFF

HAROLD P. VAN COTT, *Managing Editor*
ANITA DEVIVO, *Executive Editor and General Manager*
LINDA BARODY, *Supervising Technical Editor*
ANN G. CORBETT, *Technical Editor*
FORREST T. MULLINS, *Manager, Publications Business Service*
ANNE REDMAN, *Subscription Manager*

Published bimonthly in one volume per year, six issues (February, April, June, August, October, and December), by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D. C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D. C., and additional mailing offices.

The **Journal of Educational Psychology** publishes original investigations and theoretical papers dealing with problems of learning and teaching and with the psychological development, relationships, and adjustment of the individual. Preference is given to studies of the more complex types of behavior, especially in or relating to educational settings. Journal articles pertain to all levels of education and to all age groups.

Manuscripts: Submit manuscripts in triplicate to the Editor, Joanna Williams, P.O. Box 51, Teachers College, Columbia University, New York, New York 10027, according to instructions provided elsewhere in the Journal.

Copyright and Permissions: All rights reserved. Written permission must be obtained from the American Psychological Association for copying or reprinting text of more than 500 words, tables, or figures. Permissions are normally granted contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of $10 per page of APA material. Permission fees are waived for authors who wish to reproduce their own articles, or when material is reproduced in limited quantities for instructional purposes. Address requests for permission to the Permissions Office, American Psychological Association, 1200 Seventeenth Street, N. W., Washington, D.C. 20036.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1974: $24 domestic, $25 foreign, $5 single issue. APA member rate: $10. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive galley proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Department, APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P. O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. Undelivered copies resulting from address changes will not be replaced; second-class mail is not forwarded unless subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Date ... 8.10.74
il No. ...
Bureau Ednl Psy Research

JOURNAL OF EDUCATIONAL PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

Volume 66 June 1974 Number 3

Socialization Dimensions among Inner-City Five-Year-Olds and Later School Success: A Follow-Up 285
H. ELLISON PUSSER AND BOYD R. McCANDLESS

Reasoning and Praise: Their Effects on Academic Behavior 291
SUZANNE JOHNSON TAFFEL, K. DANIEL O'LEARY, AND SANDRA ARMEL

Individual Differences in Learning from Pictures and Words: The Development and Application of an Instrument 296
JOEL R. LEVIN, PATRICIA DIVINE-HAWKINS, STEPHEN M. KERST, AND JOSEPH GUTTMANN

The Influence of Students' Speech Characteristics on Teachers' Evaluations of Oral Answers 304
THOMAS K. CROWL AND WALTER H. MacGINITIE

Male and Female Teacher Attitudes as a Function of Students' Ascribed Motivation and Performance Levels 309
LARRY J. BRANDT AND MARY ELLEN HAYDEN

Changes in Memory Structure and Retrieval over the Course of Instruction 315
ELIZABETH F. LOFTUS AND GEOFFREY R. LOFTUS

Measuring Teacher Effects on Pupil Achievement 319
DONALD J. VELDMAN AND JERE E. BROPHY

A Follow-Up Study of Teacher Expectancy Effects 325
EDWARD J. O'CONNELL, JEROME B. DUSEK, AND RICHARD J. WHEELER

Black and White Children's Comprehension of Standard and Nonstandard English Passages 329
SAMUEL J. MARWIT AND GAIL NEUMANN

Variability in Children's Comprehension of Syntactic Structures 333
ALAN M. LESGOLD

Task, Learner, and Treatment Variables in Instructional Design 339
JOHN E. RHETTS

Teacher Behavior across Ability Groups: A Consideration of the Mediation of Pygmalion Effects 348
JUDITH LANDON ALPERT

Type and Frequency of Questions in Processing Textual Material 354
JOHN P. RICKARDS AND FRANCIS J. DI VESTA

Subject Areas and Cognitive Press 363
JOE M. STEELE, HERBERT J. WALBERG, AND ERNEST R. HOUSE

Acquisition of College Course Material under Conditions of Repeated Testing 367
JAMES H. CROUSE

Predictability of Students' Evaluations of College Teachers from Component Ratings 373
GRACE FRENCH-LAZOVIK

(Continued on next page)

Effects of Teacher Sex and Student Sex on the Evaluation of College Instructors 386
PATRICIA B. ELMORE AND KAREN A. LAPOINTE

Changing Teacher and Student Behavior: An Empirical Investigation 390
THOMAS L. GOOD AND JERE E. BROPHY

The Effects of Readability on Oral and Silent Reading Rates 406
ESTHER U. COKE

A Statistical Test of the Theoretical Model for the Representational Level of the Illinois Test of Psycholinguistic Ability 410
EUGENE B. DOUGHTIE, JAMES A. WAKEFIELD, JR., RAYMOND N. SAMPSON, AND HERBERT L. ALSTON

Some Psychometric Questions in the Evaluation of Professors 416
RICHARD M. JAEGER AND TOM D. FREIJO

Prediction of Productivity from Personality Test Scores 424
CHARLES F. ELTON AND HARRIETT A. ROSE

Classificatory Behaviors of Low-Socioeconomic-Status Children 432
GORDON K. NELSON AND HERBERT J. KLAUSMEIER

Sensitivity and Validity of Learning-Potential Measurement in Three Levels of Ability 439
ELISHA Y. BABAD AND MILTON BUDOFF

Instructional Objectives as Directions to Learners: Effect of Passage Length and Amount of Objective-Relevant Content 448
ROBERT KAPLAN AND ERNST Z. ROTHKOPF

List of Manuscripts Accepted 372

Announcement 389

SUGGESTIONS TO CONTRIBUTORS

In preparing articles for publication in the Journal, authors are strongly advised to follow the general directions given in the *Publication Manual of the American Psychological Association* (1967 Revision). All articles must be preceded by an abstract of 100–120 words. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833). Special attention should also be given to the section on the preparation of the references (*Publication Manual*, p. 39). In reference lists, give journal titles in full; do not abbreviate. *All copy must be double spaced, including the references.* All manuscripts should be submitted in *triplicate*, one of which should be an original typed copy. Since the reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the identity of the author or his affiliation should be on separate pages. Every effort should be made by the author to see that the manuscript itself contains no clues as to his identity. Dittoed and mimeographed copies are not acceptable and will not be considered. Original figures should be high-contrast, glossy, black and white prints of professional quality. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail and to check carefully the typing of the final copy.

Journal of Educational Psychology
1974, Vol. 66, No. 3, 285-290

SOCIALIZATION DIMENSIONS AMONG INNER-CITY FIVE-YEAR-OLDS AND LATER SCHOOL SUCCESS: A FOLLOW-UP[1]

H. ELLISON PUSSER[2] AND BOYD R. McCANDLESS[3]

Emory University

Five factors of socialization were originally established for a population of poor blacks and whites (86 boys and 95 girls) from the same classes: (1) verbal facility, (2) coping with anxiety by withdrawal, (3) coping with anxiety by aggression, (4) alienation, and (5) sex. These children were followed up by tests of reading readiness (kindergarten) and reading achievement (first and second grades). Significant percentages of the achievement variance were accounted for at all three grade levels by Factor 1 (except for end of first-grade males); by Factor 2 only for kindergarten girls; by Factor 3 for girls but not boys at all three grade levels; and by Factor 4 only for end of second-grade boys. Factor 5 had no statistical usefulness. Factors 1 through 4 accounted for 30% of the variance of end of second-grade boys and nearly 50% of that for end of second-grade girls. It is concluded that the usefulness of the battery for these five factors, requiring as it does only a half hour per child at the preschool level, is substantial.

In a study of economically disadvantaged, black and white children in metropolitan Atlanta begun in early 1967, Richards and McCandless factor analyzed the results from a series of reasonably reliable tests, most of which have satisfactory validity associations (See Richards & McCandless, 1972, for the data and descriptions of the tests). Five major factors in the socialization of inner-city, black and white, boys and girls were revealed. They are briefly sketched below. In the present study, the authors follow up as many of these children as could be located and make an analysis of the similarity of the present to the original sample (no real difference).

Richards and McCandless (1972) revealed five major factors in the socialization of inner-city, black and white, four- and five-year-old boys and girls.

Factor 1 was called *verbal facility*. Major loadings are for teacher ratings (for all teacher ratings, see Goldstein & Chorost, 1966) of verbal skills, quality of speech, and activity versus passivity of speech, with the addition of a loading on tested verbal intelligence.

Factor 2 was called *coping with anxiety by withdrawal*. Heaviest loadings are teacher ratings for isolation, unhappiness, fearfulness or tearfulness, "the silent child," and "the child with separation problems."

Factor 3 was called *coping with anxiety by aggression*. Highest loadings are for teacher ratings of aggressive reactions, provocativeness to teachers, disruptiveness, low cooperation, low restraint of motor activity, and hyperactivity.

Factor 4 was called *alienation*. High loadings are from the Children's Self-Social Constructs Test (Long & Henderson, 1967, 1968; Long, Henderson, & Ziller, 1970). The four subtests that contributed to this

[1] The authors appreciate the cooperation of the Atlanta Public Schools and particularly the help given by Jarvis Barnes, assistant superintendent, research and development.

[2] Now with the Psychology Department, Lander College, Greenwood, South Carolina.

[3] Requests for reprints should be addressed to Boyd R. McCandless, Psychology Department, Emory University, Atlanta, Georgia 30322.

factor were Identification with Friends, Identification with Teacher, Identification with Father, and Identification with Mother.

Factor 5 is *sex*. The major loadings are biological sex and "psychological" sex, as measured by the It Scale for Children (Brown, 1956).

These factors were derived from data collected from children attending prekindergarten educational intervention classes in 1966, 1967, and 1968, at which times they were four and five years old.

In this paper, the present authors present data about the subsequent school achievement of all of these children who could be located in the Atlanta Public Schools. Data are available for the large sample (Sample 1 of Richards and McCandless, 1972) of children who have now completed the second grade. Some data are also available for the smaller Sample 2 of Richards and McCandless. School success (as judged from standardized school readiness or achievement scores) is interrelated with the five factors of socialization summarized above.

Method

Subjects

Sample 1. From the Richards and McCandless (1972) data, complete test data were available for 181 of the Richards and McCandless Sample 1. There were 86 boys and 95 girls in this group. These children attended prekindergarten in 1968–1969. The black–white ratio was approximately 30% white to 70% black. One hundred and forty-one of these children (69 boys, 72 girls, same racial ratio) were located at the end of their kindergarten year; 125 (60 boys, 65 girls, black–white ratio of 3 to 1) were located at the end of their first-grade year; and 73 (39 boys and 34 girls) were located at the end of their second-grade year. The racial ratio continued at 3 blacks to 1 white.

Sample 2. The children from the prekindergarten classes of 1966 and 1967 are grouped together in Sample 2, which is much smaller than Sample 1. The original number was 74, with 32 boys and 42 girls and, again, there was a black–white racial ratio of 3 to 1. After the kindergarten year, only a small number of subjects could be located and follow-up multiple regression analyses were no longer justified.

All children in Samples 1 and 2 come from neighborhoods classified as poverty areas by Office of Economic Opportunity standards.

Metropolitan Readiness Tests or Metropolitan Achievement Tests were employed as the criterion of academic achievement. These tests are routinely given in April of each year at these grade levels in the Atlanta Public School system.

Factor scores from the various factor-analytic solutions of Richards and McCandless' (1972) data for subjects in Samples 1 and 2 were correlated with their Metropolitan Readiness and/or Achievement Test scores by the use of multiple, stepwise regression analyses. The analysis provided the cumulative percentage of variance for Metropolitan Readiness and Achievement Tests accounted for by each factor.[4] The purpose of these analyses was to determine the relations between the factor scores and subsequent school performance as measured by Metropolitan Readiness and/or Achievement Test scores.

Since there was such a high rate of loss of follow-up subjects from 1969 to 1972, all follow-up samples were compared to the original sample to determine if each of the follow-up samples was representative of the original. Peabody Picture Vocabulary Test scores of each of the follow-up samples were compared to the scores of the subjects in the original sample (1968–1969), and no significant differences were found for either males or females. In addition, the race and sex ratio of each of the follow-up samples did not differ significantly from the original sample. Therefore, the evidence indicates that each of the follow-up samples, although reduced in size from the original sample, is representative of the original sample.

Results

The correlations between Factors 1 to 4 and Metropolitan Reading Readiness Test or Metropolitan Achievement Test data are shown in Table 1.

Within Sample 1, whose members attended prekindergarten in the academic year 1968, the Metropolitan Readiness Test data for males reveal that there is a significant correlation (at the .01 level) between Factor 1 (verbal facility) and level of reading readiness. For female subjects, there are significant relations for the first three factors—Factor 1, verbal facility at the .01 level; Factor 2, coping with anxiety by withdrawal (the less such coping, the higher

[4] When data are analyzed by sex as they are in this paper, Factor 5 (sex) is meaningless. Factor 6 is a demographic factor of relative advantagement and enters into no meaningful relations. Factor 7 seems to represent test-specific and error variance and disappears when analysis is conducted by sex of subject. These three factors are mentioned only because they entered into the original calculations, but they are not discussed in the paper.

the readiness level at the .01 level); and Factor 3, coping with anxiety by aggression (the lower the standing in this factor, the higher the readiness level, at the .05 level). Readiness tests were administered at the end of these subjects' kindergarten year, one year after the factor measures had been made.

For the Metropolitan Achievement Test data at the end of the first grade, two years after factor assessment had been made, the following findings are revealed. For males, there are no significant correlations between achievement and any factor score. For females, there are significant correlations between achievement and scores for Factors 1 and 3, both at the .01 level. The higher the verbal facility factor score and the lower the coping with anxiety by aggression score, the higher the achievement standing.

For the Metropolitan Achievement Test data at the end of the second grade, three years after factor assessment had been made, the following significant correlations were found. For males, there are significant correlations between achievement scores and scores for Factors 1 and 4, both at the .05 level. The higher the verbal facility factor and the higher the alienation score, the higher the achievement standing. For females, there are significant correlations at the .01 level between achievement and Factors 1 (verbal facility) and 3 (coping by aggression). The higher the verbal facility factor score and the lower the coping with anxiety by aggression score, the higher the achievement standing for females.

The cumulative percentage and the levels of significance of the percentage of the variance of the Metropolitan Readiness Test and Metropolitan Achievement Test scores accounted for by the various factors are shown in Table 2. For Metropolitan Readiness data at the end of kindergarten among male subjects, only Factor 1 (verbal facility) accounts for a significant (.01) amount of the test score variance. Factors 1–4 together account for a total of 26.80% of the variance. For female subjects, Factors 1 (verbal facility), 2 (coping with anxiety by withdrawal), and 3 (coping with anxiety by aggression) each account for a significant amount of the Metropolitan Readiness Test variance (all at the .01 level or less), and Factors 1–4 together account for 50.87% of the Metropolitan Readiness Test data variance.

TABLE 1

CORRELATIONS (BY SEX) BETWEEN FACTORS 1–4 AND METROPOLITAN READINESS AND METROPOLITAN ACHIEVEMENT TESTS DATA

Sex	Factor			
	1. Verbal facility	2. Coping with anxiety by withdrawal	3. Coping with anxiety by aggression	4. Alienation
Metropolitan Readiness Test (at end of kindergarten)				
Male	.459**	.012	.086	.196
Female	.564**	.359**	.291*	.121
Metropolitan Achievement Test (at end of first grade)				
Male	.242	.024	.146	.024
Female	.346**	.198	.392**	.122
Metropolitan Achievement Test (at end of second grade)				
Male	.362*	.015	.209	.378*
Female	.552**	.221	.489**	.047

* $p < .05$.
** $p < .01$.

For Metropolitan Achievement Test data at the end of the first grade, no single factor accounts for a significant amount of variance among the male subjects, and Factors 1–4 together account for 8.95% of the total variance. For girls in the same follow-up, both Factors 1 (verbal facility) and 3 (coping with anxiety by aggression) account significantly (.01 level or less) for variance in achievement scores. Factors 1–4 together account for 29.57% of the total achievement test variance.

For Metropolitan Achievement Test data at the end of the second grade, Factors 1 (verbal facility) and 4 (alienation) each account for a significant (.05) amount of achievement test variance for males, and Factors 1–4 altogether account for 29.95% of the total achievement test variance. For

TABLE 2

PERCENTAGE (BY SEX) OF THE VARIANCE OF THE METROPOLITAN READINESS OR METROPOLITAN ACHIEVEMENT TESTS DATA ACCOUNTED FOR BY FACTORS 1–4

Sex	Factor				
	1. Verbal facility	2. Coping with anxiety by withdrawal	3. Coping with anxiety by aggression	4. Alienation	Total (1–4)
Metropolitan Readiness Test (at end of kindergarten)					
Male	21.02**	0.29	1.71	3.78	26.80
Female	31.82**	9.37**	9.55**	0.13	50.87
Metropolitan Achievement Test (at end of first grade)					
Male	5.87	0.06	2.68	0.34	8.95
Female	9.91**	3.47	15.40**	0.79	29.57
Metropolitan Achievement Test (at end of second grade)					
Male	11.26*	1.12	3.32	14.25*	29.95
Female	30.46**	0.42	15.64**	0.12	46.64

* $p < .05$.
** $p < .01$.

females, Factors 1 (verbal facility) and 3 (coping with anxiety by aggression) each account for a significant (.01) amount of achievement test variance while Factors 1–4 together account for 46.64% of the total achievement test variance.

Where N was sufficient for any meaningful analysis, the pattern of relations for the small Sample 2 was like that for Sample 1.

DISCUSSION

The most important result from the present study is that it is possible to predict the academic success of inner-city children by using the factors from Richards and McCandless (1972). The subjects' scores on certain variables when they were in preschool were related significantly to their academic success when they were in kindergarten and first and second grades. The only exception to this pattern was for the males at the end of the first grade.

The most important factor for predicting the academic success of both males and females was Factor 1, verbal facility, which is composed primarily of teacher-rated variables such as verbal skills, quality of speech, activity versus passivity of speech, and a verbal intelligence test (Peabody Picture Vocabulary Test). Verbal facility was consistently the most important factor relating to academic success of males and females. Therefore, an early education program for inner-city children should probably concentrate on developing the verbal facility of children. This factor is probably especially crucial for males, as will be discussed.

The next most important result is the differences in predictions for males and females. The factors were considerably less effective in predicting academic success for males than for females for all three years. The fact is that fewer of the original 48 variables were related to the academic success of males than females. The greatest difference in importance of variables occurred with Factor 3, coping with anxiety by aggression. Females who were rated by their teachers as being cooperative, nonaggressive, and nondisruptive were more likely to be academically successful. This factor was never related to academic success of males during the follow-up. The question that needs to be answered is why cooperativeness on the part of females, but not males, should be related to their academic success.

Another important result was the loss in prediction from the end of kindergarten to the end of first grade. Even more unusual than the loss of prediction at the end of first grade was the recovery of prediction at the end of the second grade. The total male achievement variance accounted for actually surpassed that accounted for at the end of kindergarten (29.95% at the end of second grade and 26.80% at the end of kindergarten). The total female achievement variance accounted for was almost the same as at the end of kindergarten (46.64% at the end of second grade and 50.87% at the end of kindergarten). These results are made even more mysterious by the finding that Factor 4, alienation, accounted for a

significant amount of male achievement variance at the end of second grade—the only time this factor accounted for a significant amount of variance for either males or females. First-order correlations were run between the original raw scores for males on the different variables that contributed to Factor 4, and it was determined that Children's Self-Social Constructs Test 3, identification with mother, was the single variable of Factor 4 that accounted for most of the achievement variance. The correlation was in the following direction: the more the alienation from the mother at preschool age, the more academically successful the male at the end of the second grade.

Certain questions about different variability for boys and girls arise. The differences in variability are not significant, although they are in the direction of the girls being more variable than the boys. There were no significant differences between different-raced boys and girls. The correlation coefficients clearly indicate continuity of tests from the preschool to the early elementary school years.

The recovery of prediction at the end of the second grade is most probably *not* an artifact of the reduced size of the follow-up sample at the end of second grade, since the follow-up sample does not appear to be different from the original sample on any crucial dimensions (see *Results* section for explanation). Therefore, the question remains as to why predictions are better for females than males, why these predictions drop completely out of sight for males at the end of kindergarten, and why alienation from the mother is related to academic success of males at the end of the second grade.

A possible explanation is as follows. Prekindergarten attendance for females is associated with a reduction of Factor 3, coping with anxiety by aggression, which for females is negatively related to verbal facility and, one may be sure, teacher approval. Thus, prekindergarten attendance may set girls up nicely for public school attendance. Prekindergarten (at least the prekindergartens in this study) may have the opposite effect on males. Four-year-old males are still treated as "boys will be boys, especially cute little boys." That is, they are allowed to be aggressive and even encouraged to be creative, spontaneous, and curious. Creativity, spontaneity, and curiosity for young males are likely to take the form of physical aggression. The kindergarten experience is more similar to prekindergarten than to later formal schooling, but when these children begin the traditional, passive–obedience-oriented first grade the following year, the females may be much better prepared than the males to adjust. Young males, especially young black males, from first grade on are no longer allowed to be cute, aggressive little boys and, therefore, are not able to cope with their anxiety in any way that would not affect their verbal facility. The males may have been pushed by the system and their teachers back toward passivity and withdrawal as coping techniques. As reported in Richards and McCandless (1972), being high in this factor for boys is associated with being low in verbal facility. Thus, no factor scores were related to the academic success of males at the end of the first grade. It seems fair to say that males ran into something of a brick wall when they entered the public schools. The differential male–female performance for the subjects in this study only began to show up strongly after the first grade. There was no difference in the verbal facility of the males and females at the end of prekindergarten. The passivity–obedience of the public schools also affected the academic prediction of the female, but not nearly so strongly as it affected the male. By the end of the second grade, the males had recovered from or adjusted to their "public school shock" as evidenced by the fact that Factors 1 and 4 were related significantly to their academic performance. The females also recovered from their less severe shock and Factors 1 and 3 nearly reached their original level of prediction of academic success.

The term "alienation from mother" may be unfortunate. The nature of the Children's Self-Social Constructs Test item which defines this score is such that, from the point of view of face validity, one might also say that a high-alienation-from-

mother score really means "independence from mother." The matter is one that is intriguing for further research, as the "alienation" or "intimacy" items in the Children's Self-Social Constructs Test have entered into rather strong and consistent relations within the data collected by Richards and McCandless (1972) and by the present authors. If the item taps *independence* from mother rather than *alienation* from mother, then the present results are logical enough, particularly with this sample, in which there was no father figure in nearly one half of the homes for both the black and the white children.

The fact that predictions for females are consistently better than for males supports the idea that the inner-city male is at a greater disadvantage than the inner-city female. It seems possible that inner-city females can be more successful academically by being cooperative and nice, while this is not true for inner-city males. This finding may be a result of the long-standing prejudice against the inner-city male.

A long-range value of the present study and of the Richards and McCandless (1972) study is that it is now possible to construct a test battery composed of the top loading variables of the factors that successfully predicted the academic success of the inner-city students. This test could be used as an aid in detecting those inner-city children who are likely to be successful and those who are not. Special curriculum could then be provided to help both of these groups experience sucess in school. This test could be administered by teachers in less than one half hour per child and could be a valuable aid in helping the teacher know more about the potential successes and failures of her students. This procedure could be especially beneficial for males, but this also seems to call for certain changes in the public schools. Inner-city males must be allowed some method of coping other than being pushed into passivity and obedience while they are in school.

REFERENCES

Brown, D. G. Sex-role preference in young children. *Psychological Monographs*, 1956, **70** (14, Whole No. 421).

Goldstein, K. M., & Chorost, S. B. *A preliminary evaluation of nursery school experience on the later school adjustment of culturally disadvantaged children.* (Final Report: Office of Education S-323) Staten Island, N. Y.: Wakoff Research Center, 1966.

Long, B. H., & Henderson, E. H. Social schemata of school beginners: Some demographic correlates. *Proceedings of the 75th Annual Convention of the American Psychological Association*, 1967, **2**, 329–330. (Summary)

Long, B. H., & Henderson, E. H. Self-social concepts of disadvantaged school beginners. *The Journal of Genetic Psychology*, 1968, **113**, 41–51.

Long, B. H., Henderson, E. H., & Ziller, R. C. *Manual for the Self-Social Symbols Tasks and the Children's Self-Social Constructs Tests.* Towson, Md.: Goucher College Psychology Department, 1970.

Richards, H. C., & McCandless, B. R. Socialization dimensions among five-year-old slum children. *Journal of Educational Psychology*, 1972, **63**, 44–55.

(Received September 10, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 291-295

REASONING AND PRAISE:

THEIR EFFECTS ON ACADEMIC BEHAVIOR[1]

SUZANNE JOHNSON TAFFEL, K. DANIEL O'LEARY,[2] AND SANDRA ARMEL

State University of New York at Stony Brook

Second-grade children were given reasons or praise for engaging in an academic task. The effects of reasoning and praise were then evaluated by assessing the subsequent independent maintenance of that task behavior. To insure plausibility of the reasons given, children similar to the subjects in the main study were asked why they studied arithmetic. Their reasons were then given to the experimental subjects in two separate studies. Verbalizing reasons to the child was as effective or more effective than praising him; the reasoning and praise subjects worked longer and completed more problems correctly than the control subjects. There was no additive effect of combining reasons with praise.

One objective of both parents and educators is to teach children to maintain certain important behaviors in the absence of immediate external reinforcement or punishment. Often they emphasize the importance of some activity by giving reasons to the child for engaging in the desired behavior. As Cheyne and Walters (1970) have pointed out, reasons provide at least two fairly distinct kinds of information. First, they delineate *what* behavior should be carried out, and second, they describe *why* that behavior should be performed.

While reasons, as described here, are frequently used as a means of persuading children to initiate or sustain certain activities, there are only a few experimental investigations attesting to their effectiveness. Furthermore, only one study (Staub, 1972) has investigated the effects of reasons alone rather than in conjunction with some other treatment procedure. Two studies, for example, involved reasoning in combination with punishment. In the first, Cheyne and Walters (1969) told some subjects that they were not to play with certain toys because the experimenter did not have any more like them and he was afraid that they might become worn out or broken. Other children were simply told not to touch or play with the toys. In a subsequent resistance-to-temptation session in which subjects were left alone with the prohibited toys, those children who had been given a *reason* for not playing with the toys touched them significantly less often and for shorter periods of time. A similar finding was reported by Cheyne (1969), who verbally punished some children for selecting a certain toy by saying "that is bad." Other children were also told "you should not play with that toy." Still other children were given a reason for the prohibition, "that toy belongs to someone else." When

[1] The assistance of Warren Hochberg, principal, and the second-grade teachers of Minnesauke School, Three Village School District, Stony Brook, New York, was greatly appreciated.

The project presented herein was performed pursuant to a grant from the U.S. Office of Education (OEG-0-71-2872 (607)), U.S. Department of Health, Education, and Welfare. The opinions expressed herein, however, do not necessarily reflect the position or policy of the U.S. Office of Education, and no official endorsement by the U.S. Office of Education should be inferred.

[2] Requests for reprints should be sent to K. D. O'Leary, Department of Psychology, State University of New York, Stony Brook, New York 11790.

left alone with the toy, children in the latter two conditions deviated later, less often, and for less time than children in the first condition. However, only for third-grade children was the *reason* for the prohibited behavior effective in increasing resistance to deviation. Apparently, for the kindergarten subjects involved in this study, the reason used was not capable of motivating resistance behavior either because this particular reason did not make "sense" to these children or perhaps because reasons in general are not effective with this age group. In a third investigation, by Elliott and Vasta (1970), sharing behavior was significantly increased by *adding* the reason, "if you do something nice for someone else, it means you are a good boy," to a treatment condition involving the modeling of sharing behavior with a subsequent vicarious reward. Finally, only in an experiment by Staub (1972) was it demonstrated that reasons alone could effectively increase junior high school students' choices for a larger delayed reward over a smaller immediate one.

The present investigation deviated from previous research in at least two respects. First, it employed reasons to focus on the maintenance of *academic* behavior rather than the resistance to temptation, sharing, or delay of gratification which were the target behaviors of the studies cited above. Second, unlike the Cheyne and Walters (1969), Cheyne (1969), and Elliott and Vasta (1970) experiments which used reasons *in conjunction* with punishment or modeling, the present study attempted to delineate the effects of reasons alone.

In the present study, in addition to reasoning, a second experimental group, involving praise statements for engaging in the academic target behavior, was also used. Since a number of investigations have previously demonstrated the effectiveness of praise upon the study behavior of elementary school children (Becker, Madsen, Arnold, & Thomas, 1967; Cossairt, Hall, & Hopkins, 1973; Hall, Lund, & Jackson, 1968; Madsen, Becker, & Thomas, 1968; Thomas, Becker, & Armstrong, 1968) a comparison between the praise and reasoning experimental conditions seemed to be a particularly appropriate means of measuring the relative effectiveness of reasons on an academic task; that is, reasons are being compared to a praise procedure because the latter has been shown to be effective in a host of studies.

The overall purpose of this investigation was to assess the effectiveness of reasons in children's maintenance of an academic task by comparing such a procedure to both an experimental group that was praised for academic behavior and to a control group that was given neither praise nor reasons for engaging in the academic task.

Experiment 1

Method

This study was a 2 × 2 factorial design in which some children were given reasons (reasoning condition) for engaging in the academic behavior, some were given praise (praise condition) for doing so, some were given both reasons and praise (reasoning and praise condition) and some were given neither praise nor reasons (control condition) for working at the academic task.

Subjects

The subjects were 40 second-grade children from a public elementary school, 20 males and 20 females. The subjects were white, middle-class suburban residents with average intelligence. The investigators obtained written permission from each child's parents prior to the child's involvement in the study.

Procedure

The subjects were randomly assigned to one of the four conditions with the stipulation that an equal number of males and females were involved in each condition. Each child was seen individually by the experimenter throughout the treatment or control procedure. The study was conducted in the fall of the year.

Each subject, upon entering the experimental room, was seated at a desk and given a very large booklet of grade-appropriate arithmetic problems which began with very easy problems that became progressively more difficult. The experimenter asked the subject to begin working on the problems and spent an initial 5 minutes sitting with the child as he worked. At the end of this time, the experimenter told the subject that she had some work to do at a desk in another part of the experimental room. The subject was told he could work on the problems for as long as he wished and could stop whenever he wanted to. The experimenter then returned to her desk and busied herself with "work" until the subject informed her

that he wished to stop working on the problems. At that time, the experimenter escorted the subject back to his class. If the subject did not independently stop work after 45 minutes, the experimenter terminated the session. The dependent measures used were the time the subject spent independently working on the problems and the number of problems the subject completed correctly after the experimenter had left the subject to work on his own.

During the initial five minutes of the experimental session, the 10 children in the reasoning condition were given four reasons by the experimenter for doing arithmetic. The reasons used were obtained from another group of second graders in a different public school within the same school district who answered the question, "why should we study arithmetic?" The experimenter verbalized the following reasons at one-minute intervals while the subject was working on the problems:

1. It's important to learn arithmetic so that you will know how much change you should get back when you buy something at the store. That way you won't get cheated.
2. It's important to learn arithmetic so that people will think you are smart and not dumb.
3. It's important to learn arithmetic so that you can measure the amount of flour you need to bake a cake, or so that you can measure the amount of wood you need to build a shelf.
4. It's important to learn arithmetic because you need it for lots of jobs that you might want to get when you grow up; you need arithmetic to be a scientist, a teacher, a carpenter, a mechanic, a businessman, and for many, many other jobs too.

The 10 children in the praise condition received praise instead of reasons from the experimenter during the initial five minutes of the session. The praise statements consisted of comments such as "you are doing very well," "good boy (girl)," "you're doing just fine," "good job." The comments were also given at one-minute intervals while the subject was working on the problems.

In the reasoning and praise condition, the 10 subjects were given both the praise statements and the reasons at one-minute intervals during the initial five minutes of the session.

Finally, control condition subjects were given neither praise nor reasons for working on the problems. The experimenter simply asked the subject to begin work on the problems and then sat with the subject for the initial five minutes of the session, making no comments on his performance. Control subjects, like experimental subjects, were then told they could work on the problems for as long as they wished and could stop whenever they wanted to.

It was predicted that the subjects in both the reasoning and praise conditions would work longer and complete more problems correctly than control subjects. No differential prediction was made concerning the reasoning versus the praise subjects.

Results

Due to excessively large and heterogeneous variances, Mann-Whitney U tests were executed. On number of problems completed correctly, no significant differences between any of the conditions were found. On the time data, however, the reasoning condition was significantly better than the control condition ($U = 23$, $p < .025$, one-tailed test); the praise condition was also significantly better than the control condition ($U = 18$, $p < .01$, one-tailed test), and the reasoning and praise condition was significantly better than the control condition ($U = 18.5$, $p < .01$, one-tailed test). There were no significant differences between any of the experimental groups.

Because the children were working on the problems when the reasons and praise statements were given, there was some question concerning how well the subject was attending to the experimenter's verbalizations. To insure the subject's attention to the experimenter's comments, a second experiment was performed.

Experiment 2

Method

Since there appeared to be no additive effect of combining reasons with praise in experiment 1, only three groups were used in this investigation: the reasoning condition, the praise condition, and the control condition.

Subjects

The subjects were 30 second graders, 15 males and 15 females, from the same public elementary school used in experiment 1. Parental consent was again obtained prior to each subject's involvement in the study. This study was conducted in the spring of the year.

Procedure

The general procedure remained the same as in experiment 1; however, two important changes were made. First, the experimenter no longer verbalized at one-minute intervals during the initial five minutes of the reasoning and praise conditions. Instead, the subject was given a sheet with

five easy problems on it. Upon completion of the problems, to insure the subject's attention, the experimenter removed the sheet and verbalized either a reason or a praise statement, depending upon the condition. The experimenter then handed the subject a second sheet of five easy problems. Again, upon completion of the problems, the experimenter removed the sheet and verbalized a second reason or praise statement. This procedure was continued until the subject had heard either four reasons or four praise comments from the experimenter. Control subjects were given each of the four problem sheets in succession with no verbal comments from the experimenter. All subjects were then given a large booklet of grade-appropriate arithmetic problems of increasing difficulty with instructions to work on the problems for as long as they wished, but the experimenter did not move to a desk in the experimental room to "work" as in experiment 1. Instead, the experimenter left the room entirely and asked the subject to come to the door and get the experimenter when he was finished working on the problems. It was hoped that this second procedural change—removing the experimenter from the experimental room—would reduce the amount of time the subjects voluntarily worked on the problems and, consequently, help eliminate the large and heterogeneous variances found in experiment 1.

Both the number of problems completed correctly and the time the subject spent working on the problems while he was left alone in the experimental room were used as the dependent measures.

It was again predicted that the reasoning and praise subjects would work longer and complete more problems correctly than the controls; no differential prediction was made concerning the reasoning versus praise subjects.

Results

The data did not exhibit the excessively large and heterogeneous variances found in experiment 1. Consequently, an analysis of variance was executed on both the problem and time data. The treatment effect was significant for both number of problems completed correctly ($F = 6.47$, $p < .01$) and time ($F = 7.04$, $p < .005$). Subsequently, Mann-Whitney U tests were executed on the various comparisons. On the problem data, the reasoning condition was significantly better than the control condition ($U = 16.5$, $p < .01$, one-tailed test); the praise condition was also significantly different from the control condition ($U = 26$, $p < .05$, one-tailed test); and the subjects in the reasoning condition did better than the subjects in the praise condition ($U = 27$, $p < .10$, two-tailed test). On the time data, only the reasoning condition was significantly different from the control condition ($U = 14$, $p < .01$, one-tailed test). However, there was a tendency for praise condition subjects to work longer than the controls ($U = 28$, $p < .10$, one-tailed test), and again, there was a tendency for the reasoning condition subjects to do better than the praise condition subjects ($U = 28$, $p < .12$, two-tailed test).

Discussion

If one is concerned with a child's independent maintenance of academic behavior, the results of these two experiments indicated that verbalizing reasons to a child for engaging in an academic task is as effective or more effective than giving him praise for doing so and is certainly more effective than giving him no reasons or praise statements at all. The finding that reasons are at least as effective as praise in maintaining academic behavior has obvious interesting implications for the classroom proper. However, it should be noted, that while praise has been repeatedly demonstrated to be an effective tool for modifying behavior within the classroom itself (O'Leary & O'Leary, 1972), reasons have been experimentally investigated only in the context of relatively brief single sessions. It seems important to extend investigations to repeated sessions in the natural classroom environment. A comparative study of the effects of praise and reasons in a field setting, for example, might prompt some rapprochement between two rather divergent streams of research in child psychology, namely, the operant and the cognitive approaches.

REFERENCES

Becker, W. C., Madsen, C. H., Arnold, C. R., & Thomas, D. R. The contingent use of teacher attention and praise in reducing classroom behavior problems. *Journal of Special Education*, 1967, **1**, 287–307.

Cheyne, J. A. Punishment and reasoning in the development of self-control. Paper presented at the R. H. Walters Memorial Symposium at the biennial meeting of the Society for Research in Child Development, Santa Monica, 1969.

Cheyne, J. A., & Walters, R. H. Intensity of

punishment, timing of punishment, and cognitive structure as determinants of response inhibition. *Journal of Experimental Child Psychology*, 1969, **7**, 231–244.

CHEYNE, J. A., & WALTERS, R. H. Punishment and prohibition: Some origins of self-control. In T. M. Newcomb (Ed.), *New Directions in Psychology*. New York: Holt, Rinehart & Winston, 1970.

COSSAIRT, A., HALL, R. V., & HOPKINS, B. L. The effects of experimenter's instructions, feedback, and praise on teacher praise and student attending behavior. *Journal of Applied Behavior Analysis*, 1973, **6**, 89–100.

ELLIOTT, R., & VASTA, R. The modeling of sharing: Effects associated with vicarious reinforcement, symbolization, age and generalization. *Journal of Experimental Child Psychology*, 1970, **10**, 8–15.

HALL, R. V., LUND, D., & JACKSON, D. Effects of teacher attention on study behavior. *Journal of Applied Behavior Analysis*, 1968, **1**, 1–22.

MADSEN, C. H., BECKER, W. C., & THOMAS, D. R. Rules, praise, and ignoring: Elements of elementary classroom control. *Journal of Applied Behavior Analysis*, 1968, **1**, 139–150.

O'LEARY, K. D., & O'LEARY, S. G. *Classroom management: The successful use of behavior modification*. New York: Pergamon Press, 1972.

STAUB, E. Effects of persuasion and modeling on delay of gratification. *Developmental Psychology*, 1972, **6**, 166–177.

THOMAS, D. R., BECKER, W. C., & ARMSTRONG, M. Production and elimination of disruptive classroom behavior by systematically varying teacher's behaviors. *Journal of Applied Behavior Analysis*, 1968, **1**, 35–45.

(Received September 5, 1972)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 296-303

INDIVIDUAL DIFFERENCES IN LEARNING FROM PICTURES AND WORDS:

THE DEVELOPMENT AND APPLICATION OF AN INSTRUMENT[1]

JOEL R. LEVIN,[2] PATRICIA DIVINE-HAWKINS, STEPHEN M. KERST, AND JOSEPH GUTTMANN

University of Wisconsin—Madison

An instrument was developed for determining whether or not an individual learns relatively better from pictures than words. Based on this instrument, repeated classifications of elementary school children were found to be quite consistent. Moreover, when applied to the comprehension of prose materials, the instrument served to identify those children for whom self-generated visual imagery would constitute an effective organizational strategy.

In a recent investigation, Levin, Rohwer, and Cleary (1971) demonstrated that children could be reliably classified on the basis of whether they learned relatively better from pictures as opposed to words. Following the administration of a paired-associate list that contained both unlabeled picture pairs and aurally presented word pairs, elementary school children were grouped according to whether their paired-associate recall resulted in relatively large or relatively small picture–word differences. It was found that these initial classifications of the children tended to be fairly stable over a two-day period when a second (parallel) paired-associate task was administered. More recently, Mallory (1972) has reported a similar finding based on different materials and classification procedures.

The observed stability of individual differences in the Levin et al. (1971) study was actually the serendipitous by-product of an experiment with an unrelated objective, conducted by Rohwer, Ammon, Suzuki, and Levin (1971). Thus, while the Levin et al. result turned out to be interesting in its own right, the original experiment had not been planned to answer questions directly related to individual differences. Consequently, all analyses in the Levin et al. article were admittedly post hoc, and all conclusions were admittedly speculative. Such is not the case here. Based on the previous findings, we sought to demonstrate that individual-difference-related picture–word effects obtained on a paired-associate learning task are not only reliable but also applicable to more schoollike activities such as reading.

Specifically, the dual objectives of this study were (*a*) to develop a paired-associate learning task (ideally, group administered) consisting of both pictorial and verbal items, from which different types of learners could be reliably identified, and (*b*) to determine whether such information

[1] This research was done at the Wisconsin Research and Development Center for Cognitive Learning, which is supported in part by funds from the U.S. Office of Education, Department of Health, Education, and Welfare, Center C-03, Contract OE 5-10-154. The opinions herein do not necessarily reflect the position or policy of the Office of Education and no official endorsement by the Office of Education should be inferred.

We are grateful to those who contributed to this research: Billie Albrecht, Cathy Bussey, Jim Horvitz, Karen McLellan, Vera Meyer, and Mary Quilling. We also wish to thank the staffs and students of Northside School in Sun Prairie, Wisconsin; Maywood School in Monona, Wisconsin; and Mount Horeb Elementary School in Mount Horeb, Wisconsin.

[2] Requests for reprints should be sent to Joel R. Levin, Research and Development Center, University of Wisconsin, 1025 West Johnson Street, Madison, Wisconsin 53706.

could be applied to the learning of prose materials.

Experiment 1

Method

Construction of the learning task. From a population of over 100 line drawings of objects (animate and inanimate) familiar to young children, 64 were selected to create two 16-pair lists. Only pictures with labels for which there was consensus (based on a pilot testing of first and sixth graders) were included. The 64 pictures were nonsystematically assigned to the two lists, subject to the following three restrictions: (*a*) Approximately equal numbers of animate objects were assigned to each list; (*b*) objects that were conceptually similar (e.g., knife–fork, bus–truck) were assigned to different lists; and (*c*) objects whose labels were acoustically similar (e.g., bat–cat, tire–fire) were assigned to different lists.

Within each 32-item list, the 16 pairs were nonsystematically formed subject to the following two restrictions: (*a*) Objects that were obvious associates of one another (e.g., bus–tire, doll–house) were assigned to different pairs; and (*b*) objects were paired only if there was a possibility of constructing a plausible interaction between the two.[3] Following this, 8 of the pairs in each list were randomly designated as picture pairs and 8 as word pairs. Item pairs were then randomly ordered within each list such that different item types (pictures or words) appeared in the first two list positions, as well as in the last two; in addition, no more than two consecutive pictures or words were permitted. These measures were taken as precautions against primacy, recency, and response-set effects. Four such "random" orders of the items were constructed, two for study trials and two for test trials (stimulus terms only), in order to prevent serial learning of the responses. Successive revisions and replacements of items (suggested by item analysis), as well as revisions of the instructions and procedures, were conducted to improve the task's parallel forms (separated by 24 hours) reliability. This included the addition of a third study–test cycle (actually a repetition of the first study–test trial items).

Both the pictures and typewritten labels of the pictures were photographed and mounted on slide transparencies (1 adjacent pair of pictures or words per slide). Pictures and words were placed in their predesignated positions within the 16-pair mixed-list sequences.

In the initial pilot testing of the instrument, both first and sixth graders had been tested individually. However, the final version seemed to lend itself to group administration. Accordingly, (*a*) only children in Grades 4–6 were included as subjects; and (*b*) individual subject response booklets were used on each test trial. In these booklets were printed the labels of the 16 stimulus terms, with subjects required to supply the missing response terms.

Procedure. The subjects were run in groups (intact classrooms). After distributing the response booklets, the experimenter informed the subjects of their task. Items for the first study trial of Form A were then projected onto a screen at the front of the room, five seconds per pair. Following the last study pair, subjects (having been reminded to work quickly) were allowed 1½ minutes to complete the first page of their (three-page) response booklets. Two additional study–test cycles were then provided. The next day, the experimenter returned unannounced and administered Form B of the task in similar fashion.

Subjects. Fifty-four subjects from two fourth-grade classrooms in a semirural midwestern community were administered the task. However, due to second-day absences and obvious cases of noncompliance with task instructions, only the data of 43 subjects were usable.

Results and Discussion

Subject classifications based on the learning of pictures and words. It should be recalled that a primary objective of the present experiment was to corroborate the Levin et al. (1971) finding that an individual's relative performance based on pictures and words was reliable. In the earlier study, it had been hoped that some subjects would learn pictures better than words, while the converse would be true for other subjects. However, pictures led to superior learning for almost all subjects and as a result, classification of learner types was made on the basis of a subject's picture–word difference being relatively (i.e., as compared with other subjects) large or small. As was argued by Levin et al., this kind of classification procedure created some interpretive difficulties, since with such a system a subject's performance *level* and *pattern* may be confounded. Consequently, an alternative system for classifying subjects was incorporated here.

Ideally, we hoped to identify four different types of learners: (*a*) subjects who performed relatively well on both pictures and words (high pictures, high words), (*b*) subjects who performed relatively poorly on both pictures and words (low pictures, low words); (*c*) subjects who performed rela-

[3] The second restriction was included for reasons unrelated to the experiments reported here.

tively well on pictures but relatively poorly on words (high pictures, low words); and (*d*) subjects who performed relatively poorly on pictures but relatively well on words (low pictures, high words). In actuality, however, learners of the fourth type were difficult to find on this task (only a handful out of 288 subjects in the Levin et al., 1971, study and 2 out of 43 subjects here). Fortunately even without such subjects, some interesting outcomes may be anticipated. Consider the high-picture, low-word subjects, for example. When learning pictures, their performance should resemble that of high-picture, high-word subjects rather than that of low-picture, low-word subjects; whereas when learning words, their performance should resemble that of low-picture, low-word subjects rather than that of high-picture, high-word subjects. If true, an Aptitude × Treatment interaction of the kind described by Levin (1972) would be produced, such that for some children it is largely a function of the nature of materials (here, pictures or words) that determines whether or not they will display effective learning. The stability of this type of interaction was what interested us here.

To investigate this possibility, we classified subjects according to whether or not they learned relatively well from pictures: Those who scored above the mean for pictures were designated high-picture subjects, while those below the mean were designated low-picture subjects. Within the high-picture classification, subjects were divided into two approximately equal-sized groups on the basis of their performance on words (either high word or low word). As was indicated earlier, when the same criteria for words were applied to low-picture subjects, only two subjects were found to be low-picture but high-word learners. The results of these two subjects were not included in the classification stability analysis. The number of subjects represented in the three learner classifications (high picture, high word; high picture, low word; and low picture, low word) as well as the corresponding picture and word means are shown in Table 1. Statistical analysis of Form A (the classification list) data revealed significant differences in the three groups' performance on both pictures ($F = 38.42$, $df = 2/38$, $p < .001$) and words ($F = 58.92$, $df = 2/38$, $p < .001$). The nature of these differences differed with the type of item considered, however. Consistent with the desired classification, Scheffé post hoc comparisons ($\alpha = .05$) revealed that (*a*) on pictures, both high-picture, high-word and high-picture, low-word subjects differed significantly from low-picture, low-word subjects, though not from each other, while (*b*) on words, both high-picture, low-word and low-picture, low-word subjects differed significantly from high-picture, high-word subjects, though not from each other.

TABLE 1

MEAN PERFORMANCE ON PICTURES AND WORDS BY THE THREE LEARNER CLASSIFICATIONS, EXPERIMENT 1

Measure	High picture, high word ($n = 12$)	High picture, low word ($n = 9$)	Low picture, low word ($n = 20$)
Form A			
Pictures	16.50	13.89	8.45
Words	10.25	3.44	3.50
Form B			
Pictures	15.25	14.44	9.40
Words	10.17	5.44	5.40

Criterion list performance based on initial classifications. Considering all 43 subjects (including the two low-picture, high-word subjects), the parallel forms reliability—based on total (picture plus word) scores on Form A and Form B separated by 24 hours—was found to be .76. More importantly, however, when performance on pictures and words was broken down according to the initial classification groups, essentially the same pattern was produced as with the classification list itself (cf. Form A and Form B results in Table 1). Statistically, the previously reported results were completely substantiated using Scheffé post hoc comparisons ($\alpha = .05$).

Until now we have considered only the *average* performance of subjects in the three classification groups. Of greater interest, however, is whether individual subjects who were classified in a particular way on Form A would have been classified in the

same way on a different occasion. To answer this question, we classified subjects according to their Form B performance, following the procedures used for Form A. The combined Form A – Form B classifications may be found in Table 2, which shows that 9 out of 12 (75%) high-picture, high-word subjects, 7 out of 9 (78%) high-picture, low-word subjects, and 14 out of 20 (70%) low-picture, low-word subjects were similarly classified on the two occasions. A test of the association in these data (minus the three low-picture, high-word subjects on Form B) was significant ($\chi^2 = 34.57$, $df = 4$, $p < .001$), with the degree of predictive association as reflected by the asymmetric Goodman-Kruskal measure, λ_B (Hays, 1963), being .62.

TABLE 2

CORRESPONDENCE BETWEEN FORM A AND FORM B SUBJECT CLASSIFICATIONS

Form B	Form A		
	1	2	3
1. High picture, high word	9	1	1
2. High picture, low word	1	7	2
3. Low picture, low word	2	1	14

Note. Three subjects who were low-picture, low-word learners on Form A would have been classified as low-picture, high-word learners on Form B.

Rejecting an alternative explanation. The data just reported, as well as those of Levin et al. (1971), would seem to support the argument that different "learner types" exist and that they can be identified. However, because of the unique mixed-list nature of the learning task employed in these experiments, an alternative interpretation seems reasonable. Rather than explaining the results in terms of individual-difference-related learner types, it is possible to explain them in terms of individual-difference-related "learning styles." In particular, when both pictures and words are presented in a single list, the contrast produced enables subjects to attend to one class of materials (pictures) at the expense of the other (words). Consequently, the high picture learning of high-picture, low-word subjects may have resulted simply from an unequal division of attention between the two classes of material. A second experiment was conducted to evaluate this possibility.

The subjects consisted of sixth graders from a semirural midwestern community. Following the procedures of Experiment 1, we administered Form A of the learning task as a 16-pair mixed list which contained both picture and word pairs. On the basis of the children's performance, we were able, as previously, to identify high-picture, high-word subjects, low-picture, low-word subjects, and high-picture, low-word subjects. As expected, even though the performance of these latter children was closer to that of high-picture, high-word subjects on the picture pairs, it was closer to that of low-picture, low-word subjects on the word pairs.

On the following day, the children were administered a second learning task which differed from the previous second task in one important respect: Whereas Experiment 1 utilized a parallel form of the mixed-list classification instrument, here we constructed homogeneous (picture or word) lists incorporating the same paired members. Thus, about half of the children (two classrooms) received a list of 16 picture pairs on the second day, and the others (two classrooms) received a complementary list of 16 word pairs. If the alternative (learning style) interpretation of the previous data is valid, then with no opportunity for high-picture, low-word subjects to divide their attention unequally between pictures and words (since subjects received a list containing only one class of materials), different performance profiles would not be expected on the two list types. On the other hand, our original (learner type) interpretation would have high-picture, low-word subjects performing either well or poorly, depending on whether they were administered the picture or the word list, respectively.

As is evidenced in Figure 1, when word pairs were presented, both high-picture, low-word and low-picture, low-word subjects did very poorly; however, when picture pairs were presented, the performance of high-picture, low-word subjects improved

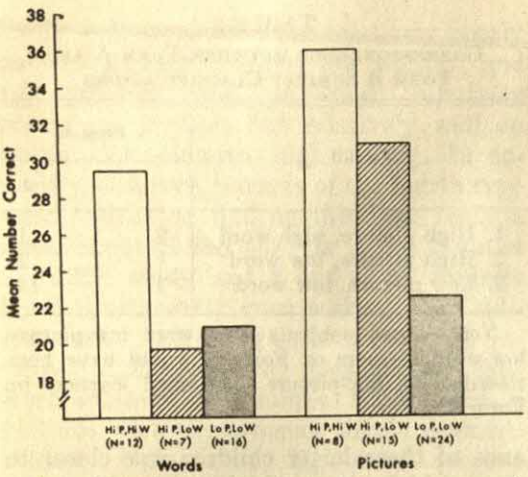


FIGURE 1. Mean performance (maximum = 48) by the three learner types on the homogeneous word and picture lists of the second day. (Abbreviations: Hi P = high picture; Hi W = high word; Lo P = low picture; and Lo W = low word.)

dramatically while the performance of low-picture, low-word subjects was not materially affected. Thus, given that profiles paralleling those of Experiment 1 were obtained with independent homogeneous lists, a "learning style" interpretation of those data must be rejected. Until proven otherwise, our "learner type" interpretation still lives—a conclusion further supported by the data of Experiment 2.

EXPERIMENT 2

A second (equally important) concern of the present research was that the previously described learner type classifications would relate to performance in learning tasks other than parallel versions of the paired-associate classification task. In particular, reading comprehension was selected as a likely candidate since it has been argued that (a) experimenter manipulations seem to affect paired-associate learning and reading comprehension in similar ways, and (*b*) similar processes may well underlie each (Levin, 1972).

For example, it is well documented that in a paired-associate task, picture pairs are more easily learned than word pairs (cf. Reese, 1970). It has likewise been shown that a pictorial representation of textlike materials is more easily learned than is a printed representation of the same materials (Matz & Rohwer, 1971). The same comparison may be made with regard to the role of subject-generated visual imagery in paired-associate learning and in reading comprehension. That is, with relatively concrete materials, the generation of imagined visual relationships has been found to facilitate both types of performance (Levin, 1972).

In a recent experiment, Levin (1973) demonstrated that while subject-generated visual imagery improves reading comprehension in general, the effectiveness of such a strategy depends largely upon the prerequisite skills of the student. Specifically, fourth graders who could decode and derive meaning from individual words (but could not effectively organize words to derive meaning from sentences) benefited greatly from instructions to generate organizational images on a reading task. As was predicted, however, children who were experiencing decoding and/or vocabulary problems at the word level did not benefit from such an imagery strategy.

An analogy might be drawn vis-à-vis the focus of the present research. Suppose that children are classified according to the system in Experiment 1, in which high-picture, low-word subjects were those children who learned relatively well from pictures but not from words. An intriguing possibility is that their comprehension of textlike (verbal) materials might be improved through the substitution or addition of pictures. On the other hand, this would not be expected for low-picture, low-word subjects who have difficulty learning from pictures as well as from words.

In this experiment, we wanted to see if the three learner type classifications differed with respect to reading comprehension under naturally occurring situations (i.e., in the absence of experimenter-suggested strategies). In addition, however, some of the subjects from each classification group were instructed to employ a visual imagery strategy while reading, with the expectation that only the performance of those subjects who learn relatively well from pictures (that is, high-picture, high-word and high-

picture, low-word subjects but not low-picture, low-word subjects) would be enhanced.

Method

Reading task. Two 10-sentence reading passages appropriate for children of ages 9–12 were constructed following Matz and Rohwer (1971) and Levin (1973). The two passages (one comparing two kinds of monkeys and the other, two cars) had been used in research recently reported by Levin and Divine-Hawkins (in press). Each sentence was photographed and mounted on slide transparencies, 1 to a slide. Ten questions based on each passage were constructed to assess comprehension.

Subjects. Children from three fourth-grade classrooms in a middle-class midwestern community participated in the experiment.

Procedure. Form A of the Experiment 1 group-administered learning task was presented to children in each classroom, following the procedures previously described. The next day subjects were called out of their rooms individually and were given the two reading passages. Additionally, half of the subjects were given a visual imagery strategy prior to reading the passages. That is, they were told to make up pictures in their minds about what was happening in each story while they read it. The subjects were then provided with a sample sentence (with subjects in the imagery condition given practice in generating images), followed by an oral question about it. The first passage was presented on a slide projector, one sentence every eight seconds. Following the last sentence, the experimenter asked the 10 questions about the passage in a random order (i.e., the questions, which were typed on index cards, were shuffled anew for each subject). No reading was required of the subject during these oral questions, each of which could be answered in short phrases. The second passage and corresponding questions were then presented in similar fashion. After the second set of questions, the experimenter queried the subject regarding his perceived passage difficulty and his interest in the two passages. The subject was also asked to indicate how frequently visual images came to mind while he was reading the passages. Four-point ordinal scales were used to quantify subjects' responses to each question.[4]

Results

Subject classifications on the learning task paralleled those of Experiment 1 and resulted in the identification of 24 high-picture, high-word subjects, 13 high-picture, low-word subjects, and 20 low-picture, low-word subjects. The mean learning of pictures and words by subjects in these three groups is presented in Table 3. As was true in Experiment 1, these classifications resulted in comparable performance for high-picture, low-word subjects and high-picture, high-word subjects on pictures and for high-picture, low-word subjects and low-picture, low-word subjects on words.

TABLE 3

MEAN PERFORMANCE ON PICTURES AND WORDS BY THE THREE LEARNER CLASSIFICATIONS, EXPERIMENT 2

Measure	High picture, high word (n = 24)	High picture, low word (n = 13)	Low picture, low word (n = 20)
Form A			
Pictures	15.58	15.00	7.25
Words	12.75	4.15	3.05

Since the experimenter assigned the subjects randomly to the two reading conditions without knowledge of their particular learner type classifications, disproportionate numbers of subjects ended up in the two conditions from one learner type to the next, as indicated in Figure 2. In scoring the reading performance data, nothing more than a synonymic deviation from the correct response was accepted. Analysis of the data (which represent the mean number of

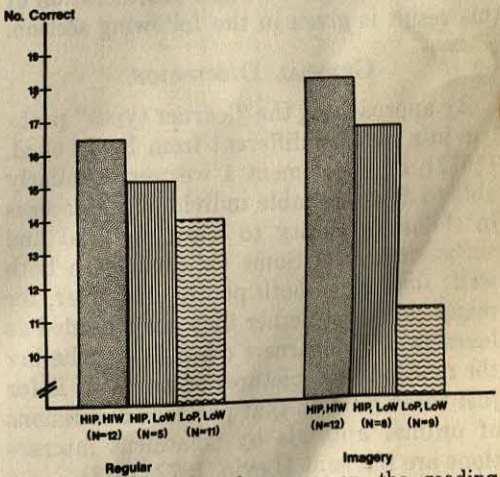


FIGURE 2. Mean performance on the reading task by the three learner types under different instructional conditions. (Abbreviations: Hi P = high picture; Hi W = high word; Lo P = low picture; and Lo W = low word.)

[4] Unfortunately, the data derived from these questions were uninformative and, therefore, are not discussed further.

correct responses, out of 20, on the two passages) in Figure 2 was performed using least squares techniques for the effects of interest. In order to compare the reading performance of the three learner types under each instructional condition (regular and imagery), learner types were nested within these two conditions.

As may be seen in Figure 2, for the subjects given regular instructions prior to reading the passages, differences among learner types were small and statistically nonsignificant ($F = 1.81$, $df = 2/51$, $p > .10$). However, when imagery instructions were employed, significant performance differences among learner types were detected ($F = 135.49$, $df = 2/51$, $p < .001$). Scheffé post hoc comparisons ($\alpha = .05$) confirmed the visual impression obtained from Figure 1: High-picture, high-word subjects and high-picture, low-word subjects each differed significantly from low-picture, low-word subjects, though not from each other.

As a main effect, imagery instructions were not facilitative ($F < 1$), the explanation of which may be inferred from Figure 2: While the performance of good picture learners (high-picture, high-word subjects and high-picture, low-word subjects) increased when imagery instructions were employed, the performance of poor picture learners declined. A further consideration of this result is given in the following section.

General Discussion

By approaching the "learner types" problem in a manner different from Levin et al. (1971), in Experiment 1 we were similarly able to detect reliable individual differences in children's ability to learn pictorial and verbal materials. Some children learn both well; some learn both poorly. However, for many children whether they are regarded as learners or nonlearners depends on whether the materials are pictures or words. It is for just these children that previous discussions of ordinal aptitude by treatment interactions are relevant (Levin, 1972, 1973).

Psychometrically speaking, it is important to note that the Experiment 1 classifications were sufficiently potent to overcome the counteracting influences of statistical regression (on Form B). Practically speaking, it is also important that learner type diagnoses may be couched within a group-administered task. At the same time, one should not lose sight of the fact that the Experiment 1 data are based on only a 24-hour separation. It would certainly be fruitful to determine the limits of the instrument's long-term stability.

In Experiment 2, we capitalized on the learner type classifications to assess a child's performance on a reading task. While minimal differences among groups were discovered on the reading task per se, when a visual imagery strategy was induced in the children prior to reading, substantial differences among learner types were observed. What we found was that children who do not learn appreciably better from pictures than from words (low-picture, low-word subjects) did not benefit as much from the imagery strategy as did those who do (high-picture, low-word subjects). In fact, as Figure 2 suggests, imagery instructions may well have been detrimental to the reading comprehension of low-picture, low-word subjects. Assuming that such subjects have developed alternative (nonimagery) strategies for successfully processing prose materials under natural conditions (cf. the bars to the left in Figure 2), this result is not totally surprising.

Just as it has been previously demonstrated that children first must comprehend individual words before they can use visual imagery to their advantage while reading (Levin, 1973), the present research adds to this finding by suggesting that certain Learning Modality × Reading Strategy interactions may also have to be considered. Of late, visual imagery has been heralded as an effective organizational strategy for relatively concrete prose materials (Levin, 1972). However, when its success clearly depends on the capabilities of the user, caveats about its nonuniversality cannot be echoed too loudly.

REFERENCES

Hays, W. L. *Statistics for psychologists.* New York: Holt, 1963.

LEVIN, J. R. Comprehending what we read: An outsider looks in. *Journal of Reading Behavior*, 1972, **4**, 18–28.

LEVIN, J. R. Inducing comprehension in poor readers: A test of a recent model. *Journal of Educational Psychology*, 1973, **65**, 19–24.

LEVIN, J. R., & DIVINE-HAWKINS, P. Visual imagery as a prose-learning process. *Journal of Reading Behavior*, in press.

LEVIN, J. R., ROHWER, W. D., JR., & CLEARY, T. A. Individual differences in the learning of verbally and pictorially presented paired associates. *American Educational Research Journal*, 1971, **8**, 11–26.

MALLORY, W. A. Abilities and developmental changes in elaborative strategies in paired-associate learning of young children. *Journal of Educational Psychology*, 1972, **63**, 202–217.

MATZ, R., & ROHWER, W. D., JR. Visual elaboration of text. Paper presented at the annual meeting of the American Educational Research Association, New York, March 1971.

REESE, H. W. (Chm.) Imagery in children's learning: A symposium. *Psychological Bulletin*, 1970, **73**, 383–421.

ROHWER, W. D., JR., AMMON, M. S., SUZUKI, N., & LEVIN, J. R. Population differences and learning proficiency. *Journal of Educational Psychology*, 1971, **62**, 1–14.

(Received March 23, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 304-308

THE INFLUENCE OF STUDENTS' SPEECH CHARACTERISTICS ON TEACHERS' EVALUATIONS OF ORAL ANSWERS

THOMAS K. CROWL[1]

Richmond College
City University of New York

WALTER H. MacGINITIE

Teachers College
Columbia University

Tape recordings were made of six white and six black ninth-grade boys speaking identically worded answers to typical school questions. Significantly higher grades were assigned by 62 experienced white teachers to the recorded answers when spoken by white students than when spoken by black students. Teachers who were most susceptible to vocal stereotyping could not be differentiated on the basis of sex, age, years of teaching experience, most frequently taught grade level, or the percentage of black students most frequently taught.

The importance of vocal cues as a basis for forming attitudinal judgments about a speaker has been dramatically demonstrated in a series of studies in which listeners who heard tape recordings of bilingual or bidialectal speakers formed significantly different attitudes toward the *same* speaker, depending on what language or dialect the speaker was speaking (Anisfeld, Bogo, & Lambert, 1962; Anisfeld & Lambert, 1964; Lambert, Anisfeld, & Yeni-Komshian, 1965; Lambert, Frankel, & Tucker, 1966; Lambert, Hodgson, Gardner, & Fillenbaum, 1960).

In studies involving listeners' judgments about speakers, it has frequently been found that listeners tend to agree with one another with respect to the judgments they make, not only in cases where these judgments are accurate but also in cases where these judgments are inaccurate. (See Kramer, 1963; Sanford, 1942, for comprehensive reviews.) Agreement among listeners with respect to the errors they make in judging a speaker's characteristics is probably the most forceful evidence of vocal stereotyping.

Williams, Whitehead, and Miller (1972) recently investigated the concept of vocal stereotyping within an educational context by examining relationships between teachers' expectations of student academic performance and teachers' attitudes toward students' speech. Using students from three ethnic groups, the study found that students whose speech was judged to be more nonstandard were also expected to perform worse academically than students whose speech was judged to be more standard.

It is important to note that the study referred to above focused on expectations of student performance. The present investigation was designed to find out if students' actual (not expected) academic performance is judged differently by teachers as a function of students' speech characteristics. An attempt was made to examine some important practical ramifications of the phenomenon of vocal stereotyping. This was done by changing the level of conceptual focus from the listeners' attitudinal judgments and expectations about a speaker to the level of overt behavior actually exhibited by listeners toward a speaker as a function of the speaker's vocal characteristics. The present study was directed at two general questions: Do differences in students' speech characteristics lead to differential judgments by teachers about students' academic performance? If so, is it possible to identify specific characteristics of teachers who are most susceptible to vocal stereotyping?

[1] Requests for reprints should be sent to Thomas K. Crowl, Professional Studies, Richmond College, City University of New York, 130 Stuyvesant Place, Staten Island, New York 10301.

Method

Selection of Speech Sample

Speech samples of two groups of ninth-grade boys were used in the study, one group consisting of six white students of an upper-middle socioeconomic background and the other group consisting of six black students of a lower socioeconomic background. It was felt that only after evidence had been gathered which demonstrated differential judgments on the basis of rather gross speech differences would it be feasible to try to isolate specific speech characteristics associated with differential teacher judgments.

In order to control the content of students' answers while varying the speech characteristics of the students, 12 predetermined answers were used for each of two questions:

> Why do we celebrate Thanksgiving?
> What is the difference between a discovery and an invention?

The predetermined answers were based on answers actually given by another group of white ninth graders of upper-middle socioeconomic level. The answers were worded in standard English, with the mean number of words being 18.3. Tape recordings were made of all students in the study speaking all answers.

To insure that each student's ethnic group could be accurately identified from his speech, 12 judges drawn from the same population of teachers who participated in the study listened to the tapes and were asked to identify each student's ethnic group. The overall accuracy of individual judges' identifications was high, ranging from 75% to 100%, with a mean of 88.4%.

Students were provided with a typed set of the predetermined answers and were given an unlimited amount of time to rehearse the answers and to ask the experimenter about the meaning or pronunciation of words or about the general procedures for taping. Answers were taped as often as necessary until both the experimenter and the student agreed that each written answer had been spoken verbatim and in a natural-sounding manner.

Preparation of Tapes

By splicing together segments from the original recordings, two tapes were compiled in such a way that on each tape (*a*) each of the 24 answers occurred once, (*b*) half of the answers to each question were given by black students and half by white students, (*c*) each student gave one answer to each question, and (*d*) for any given answer, the ethnic group of the respondent on one tape was reversed on the other tape. Answers to the Thanksgiving question preceded answers to the "discovery" question on Tape A; on Tape B, this order was reversed. The beginning of each tape contained a set of instructions and four answers spoken by the experimenter to one practice question.

Subjects

The subjects consisted of 62 white teachers who had one or more years' teaching experience and who volunteered to participate. Overall, there were more female than male subjects ($n = 35$ and 27, respectively); the modal age was 25–34 years ($n = 31$); the modal number of years teaching experience was 1–5 years ($n = 27$); the most frequent grade level taught was either elementary ($n = 30$) or senior high school ($n = 25$); the highest academic degree held was either a bachelor's degree ($n = 25$) or a master's degree ($n = 36$); and classes most frequently taught by the subjects were comprised of 75%–100% white students ($n = 45$).

Experimental Procedure

The study was presented in the guise of a project for gathering data concerning grading practices of experienced teachers in order to establish norms for grading oral answers. Teachers were instructed to grade each answer in terms of "how well it really answers the question," using a scale where 10 = excellent and 1 = completely wrong. Recordings were presented to subjects either individually or in small groups, with half of the subjects randomly assigned to listen to Tape A and half to Tape B. After grading the answers, each subject filled out an anonymous biographical data sheet. In terms of sex, age, total years teaching experience, most frequent grade level taught, highest academic degree held, and ethnic group of students most frequently taught, the subjects in both groups were remarkably similar.

Results

By summing the ratings given by the subjects across both tapes, it was possible to compare the mean rating for answers when spoken by black students with the mean rating for the same answers when spoken by white students. An analysis of variance with repeated measures was initially carried out using the ratings of answers to each question separately. Since the results of the separate analyses were virtually identical, the data were combined and reanalyzed. In all analyses, the pooled interaction of individual subjects and individual students was used as the error term. The difference between subjects' ratings of the same answers spoken by

black students ($M = 5.48$, $SD = 2.60$) and by white students ($M = 5.82$, $SD = 2.59$) was significant ($F = 13.22$, $df = 1/600$, $p < .001$). Thus, it was concluded that white teachers were influenced in their evaluations of students' oral answers by the speech characteristics of students whose ethnic group could be identified from their speech.

The finding of a significant difference between the ratings assigned by the two groups of subjects ($F = 21.74$, $df = 1/600$, $p < .001$) was unanticipated in view of the marked similarity of the two teacher groups in terms of biographical data and was particularly surprising in view of the comparability of the two groups' ratings of the practice answers. Also unanticipated was the significant interaction between the ethnic group of student and of teacher group ($F = 14.19$, $df = 1/600$, $p < .001$), indicating that subjects' ratings of answers with respect to the students' ethnic group differed, depending on which tape had been listened to by the subject. Although the same students and the same answers appeared on both tapes, answers assigned to black students on Tape A were assigned to white students on Tape B, and vice versa. It had been assumed that some of the predetermined answers were inherently better answers than others, but it had also been assumed that random assignment would result in the answers on each tape being similar in quality for the two ethnic groups. In light of the above findings, the assumption of similar distribution of answer quality was tested.

In order to examine the effect of answer quality without the effect of students' voices, written versions of the answers used in the study were presented to 42 student teachers. An analysis of the mean ratings of the answers in written form indicated that the answers assigned to white students on Tape A ($M = 6.10$, $SD = 2.80$) received significantly higher ratings ($t = 4.42$, $p < .001$) than did answers assigned to black students on Tape A ($M = 5.62$, $SD = 2.81$). Since the ethnic group of the student giving a particular answer was reversed on the two tapes, the above finding also indicated that the answers assigned to black students on Tape B received significantly higher ratings in their written form than did answers assigned to white students on the same tape.

If the inherent quality of the answer was the only factor affecting subjects' ratings of oral answers, one would expect white students' answers to receive higher ratings on Tape A and black students' answers to receive higher ratings on Tape B. In fact, white students' answers did receive higher ratings on Tape A (White, $M = 6.22$, $SD = 2.72$; Black, $M = 5.52$, $SD = 2.66$). But black students' answers did *not* receive higher ratings on Tape B (Black, $M = 5.44$, $SD = 2.54$; White, $M = 5.42$, $SD = 2.40$). In view of the evidence cited above, it appears that inherently superior answers spoken by black students were not perceived as any better than inherently inferior answers spoken by white students; or conversely, inherently inferior answers spoken by white students were perceived as being as good as inherently superior answers spoken by black students.

Since differences in students' speech characteristics led to significantly different judgments of the quality of students' academic performance, an attempt was made to identify specific characteristics of those teacher subjects who were most susceptible to vocal stereotyping. No significant differences were found, however, between ratings assigned to answers spoken by students of different ethnic groups by male or female subjects, by subjects younger than or older than 35 years of age, by subjects with less than or more than five years teaching experience, by subjects who were elementary school teachers or nonelementary school teachers, or by subjects having taught classes comprised of less than or more than 25% black students.

Discussion

In most previous studies of vocal stereotyping, subjects have been explicitly instructed to judge a speaker's characteristics on the basis of vocal cues, and it has usually been tacitly assumed that the other kinds of behavior of a listener toward a speaker would vary as a function of the way in

which the listener perceived the speaker. The current study did not make the task of judging the speaker's characteristics explicit. On the contrary, subjects were specifically instructed to evaluate the content of the speech and implicitly asked to ignore the characteristics of the speaker. A different kind of behavior—namely, evaluating answers—was measured directly, and it was inferred that differences between evaluations of answers containing the same words were a function of the listener's perception of the speaker. The overall findings of the study supported the notion that the content of the same oral answer is evaluated differently when spoken by different persons whose difference in ethnic group is identifiable from their speech.

The results of the current study are consistent with many previous findings regarding vocal stereotyping. In this study, listeners judged the content of answers spoken by black students as inferior to the content of answers spoken by white students. Such a judgment is, by definition, erroneous since the verbal content of black students' and white students' answers was the same. The results also support the findings from studies of bilingual and bidialectal speakers (Lambert et al., 1960, 1965, 1966), inasmuch as listeners of a majority ethnic group reacted unfavorably toward members of a minority ethnic group. Although no measures of subjects' attitudes toward blacks in general were obtained, it is unlikely that negative attitudes would have been expressed. The subjects participating in the study, who were experienced teachers attending graduate courses during the summer, probably comprised a group with rather liberal views. It should also be noted that subjects participating in the study came from a total of 22 different states, with only one subject from a southern state.

Some differences in vocal cues between the two sets of answers may have been directly associated not with the different ethnic and socioeconomic backgrounds of students used in the study but rather with the experimental procedure itself. For example, the fact that students rehearsed their answers might have added an automatic quality to the recordings, although anecdotal evidence from subjects indicated that teachers were not aware that the answers were not spoken spontaneously. Also, while the wording of answers in standard English was useful in eliminating rating biases associated with nonstandard grammar, it is possible that the recorded answers given by students who may characteristically speak nonstandard English may have sounded incongruent. Again, however, anecdotal evidence suggested that this possibility was somewhat remote since most teachers indicated that during the experiment they were not even aware that the students came from different ethnic and socioeconomic backgrounds.

Within an educational context, the findings of the study are consistent both with the recent finding that students with more nonstandard speech are expected to perform worse academically than students whose speech is judged as more standard (Williams, et al., 1972) as well as with the much earlier finding of Michael and Crawford (1927) that ratings of students' speech are positively correlated with students' academic performance.

While the findings of the study have implications concerning the nature of teacher–student interaction in the classroom and should alert teachers to a possible source of bias that might go undetected, it must be remembered that the experimental conditions of the study differed in a number of obvious and important ways from a typical classroom situation. In evaluating oral answers in a classroom, teachers have visual as well as vocal cues on which to base their judgments, students' responses are spoken spontaneously rather than after practice in saying them naturally, and the teacher often is already well acquainted with the student who is giving an answer. Also, controlling the content of answers may have diminished differences that might actually be found among students such as those used in the study, because vocabulary and grammatical differences were purposely eliminated. It is difficult to know how the combination of cues transmitted from the student to the

teacher might detract from an unbiased appraisal of a student's academic performance.

An important way in which the experimental conditions probably did correspond to a classroom situation is the fact that the teachers' attention was not drawn specifically to the speech characteristics of students. Despite the fact that students' speech differences were not highlighted, vocal stereotyping occurred. An immediate practical concern emanating from the finding of vocal stereotyping are the grades actually assigned to students and the impact these grades may have on the student's perception of himself and his subsequent attitudes toward school and academic performance. Although statistically significant, the magnitude of the difference between ratings assigned to speakers of different ethnic groups was only .34 on a 10-point scale. Nevertheless, if teachers engage in vocal stereotyping without being aware that they are stereotyping, the potential cumulative influence of a student's speech on the teacher's judgment of the student could be considerable, particularly when one considers the frequency of student–teacher verbal interaction in the classroom. Even though teachers may not be recording marks in a grade book every time a student speaks, it may be that the teacher makes some kind of judgment about what kind of person the student is, and these subtle judgments may ultimately affect the teacher's behavior toward a student.

It should be noted that it was not possible to identify specific characteristics of teachers who were most susceptible to vocal stereotyping, but all of the teachers in the study were white and very few of them had had much experience teaching black students. It should also be noted that the lower scores assigned to black students' answers may reflect, at least in part, differences in the perceptual ability of white teachers to apprehend black students' speech. The result of not being able to apprehend as easily what a student is saying might lead to the assignment of a lower score. Certainly, it would seem fruitful to carry out similar kinds of investigations with different populations, such as black teachers or teachers from various geographic regions or dialect areas.

REFERENCES

ANISFELD, E., & LAMBERT, W. E. Evaluational reactions of bilingual and monolingual children to spoken languages. *Journal of Abnormal and Social Psychology*, 1964, **69**, 89–97.

ANISFELD, M., BOGO, N., & LAMBERT, W. E. Evaluational reactions to accented English speech. *Journal of Abnormal and Social Psychology*, 1962, **65**, 223–231.

KRAMER, E. Judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychological Bulletin*, 1963, **60**, 408–420.

LAMBERT, W. E., ANISFELD, M., & YENI-KOMSHIAN, G. Evaluational reactions of Jewish and Arab adolescents to dialect and language variations. *Journal of Personality and Social Psychology*, 1965, **2**, 84–90.

LAMBERT, W. E., FRANKEL, H., & TUCKER, G. R. Judging personality through speech: A French-Canadian example. *Journal of Communication*, 1966, **16**, 305–321.

LAMBERT, W. E., HODGSON, R. C., GARDNER, R. C., & FILLENBAUM, S. Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology*, 1960, **60**, 44–51.

MICHAEL, W., & CRAWFORD, C. C. An experiment in judging intelligence by voice. *Journal of Educational Psychology*, 1927, **18**, 107–114.

SANFORD, F. H. Speech and personality. *Psychological Bulletin*, 1942, **39**, 811–845.

WILLIAMS, F., WHITEHEAD, J. L., & MILLER, L. M. Relations between language attitudes and teacher expectancy. *American Educational Research Journal*, 1972, **9**, 263–277.

(Received June 11, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 309-314

MALE AND FEMALE TEACHER ATTITUDES AS A FUNCTION OF STUDENTS' ASCRIBED MOTIVATION AND PERFORMANCE LEVELS[1]

LARRY J. BRANDT AND MARY ELLEN HAYDEN[2]

University of Houston

The attitudes of male and female college students were compared after having taught a successful or unsuccessful simulated student who was labeled as either an overachiever or underachiever. The data showed the following: performance of the child was found to be the predominant factor in determining the teachers' attitudes; ascriptions modified the performance effect; and some differences in attitudes were found between the male and female subjects. Implications of the data were discussed.

The manner in which teachers form their opinions of students and the effect these opinions have on the students have become important issues in educational research. Rosenthal and Jacobson (1968) initiated much of the recent work in this area with their book *Pygmalion in the Classroom.* They found that randomly labeling some children "late bloomers" at the beginning of the school year had a significant effect on those children's IQ scores at the end of the school year. The authors assumed that the increased scores reflected differential expectations and behaviors on the part of the teachers.

Subsequent experiments have yielded mixed results. Baker and Crist (1971) summarized nine unsuccessful attempts to replicate the *Pygmalion* results. However, several other studies have shown that teachers' expectations did indeed affect their teaching performance. Beez (1968) found teachers taught more to students randomly labeled "high ability" than to those labeled "low ability." Medinnis and Unruh (1971) discovered a higher positive/negative reinforcement ratio from teachers for students randomly rated as "high ability" than for those rated as "low ability." Rothbart, Dalton, and Barrett (1971) found that teachers were more attentive toward children randomly labeled "bright" than toward those labeled "dull." They also rated the "bright" students as being more intelligent, having greater potential for future success, and having less need for approval. Brophy and Good (1972) have suggested two reasons for these discrepant results. First, results inconsistent with those of Rosenthal and Jacobson (1968) generally spanned a long period of time. Second, teachers were often familiar with the students before ascriptions were made about them. Brophy and Good suggested that in this type of situation, information from an experimenter is not a very strong research manipulation; the teachers may simply disbelieve or otherwise reject the experimenter's information.

Furthermore, the dependent variable in the Rosenthal and Jacobson (1968) study and in the nine attempted replications reviewed by Baker and Crist (1971) was IQ scores. In most of the related studies investigating teacher expectancies, the dependent variables were teacher and/or student behaviors.

Both the successful and the unsuccessful studies referred to above used a paradigm that involved (*a*) making differential ascriptions about students assigned randomly

[1] This article is based on a dissertation in partial fulfillment of the requirements for the PhD degree at the University of Arizona.

[2] Request for reprints should be sent to Mary Ellen Hayden, Department of Psychology, University of Houston, Houston, Texas 77004.

from the same population, (*b*) giving the teacher the opportunity to interact with and to influence the students, and (*c*) assessing differential teacher behavior and/or the resultant student behavior. The present study involved a different paradigm in that differential ascriptions were made as in the Rosenthal and Jacobson (1968) model, but by simulating fictitious student responses, the teacher had no opportunity to affect the behavior of the child. The teacher's attitudes toward the child and toward herself were assessed at the end of the teaching situation.

It was expected that the second model would not only confirm previous findings that ascriptions affect teacher attitudes but would also show what happens when a teacher has an expectation about the student which is either confirmed or countered by his behavior. Of particular interest was whether the resultant attitude of a teacher is more influenced by the teacher's initial perception of the student or by the student's actual performance.

Other than the results recently reported by Good, Sikes, and Brophy (1973) which found that male and female teachers behaved differently in certain teacher–student interactions, the question of differential behaviors and attitudes of male and female teachers has not been studied. Thus, a second purpose of this study was to determine whether male and female subjects form different attitudes in a teaching situation.

Method

Subjects

Forty-eight subjects, half male and half female, were chosen from introductory psychology courses at a large university. Each volunteer was between 18 and 24 years of age and had no children in his immediate family under the age of nine.

Overview of Task

Each subject was informed that he would teach a six-year-old male child a discrimination learning task but that because various characteristics in addition to a child's performance (e.g., his appearance) may affect teacher–child interactions, he would neither see, hear, nor talk with the child. On a display panel, the subject saw which stimulus was presented to the child and which response was made by the child. Subsequently, the subject was free to administer feedback, additional information, and reinforcement by pressing specific panels of the apparatus. He then pressed another panel which initiated the next trial. Supposedly, an assistant was in the room with the child for the purpose of conveying information to him from the subject. In reality, there was neither an assistant nor a child present; programmed sequences of correct and incorrect responses simulated the performance of a successful or unsuccessful child. The responses of the adult, along with four rating scales completed by the subject upon termination of the teaching task, comprised the dependent variables. Only the rating scale data are reported in this paper.

A complete description of the experimental apparatus is described by Brandt (1971).

Procedure

The subjects were assigned randomly to the eight treatment combinations of a $2 \times 2 \times 2$ factorial design with the restriction that there were the same number of education and noneducation majors in each group of female subjects and that all male subjects were noneducation majors. The independent variables were (*a*) performance of the child: successful–unsuccessful; (*b*) information given to the subject about the motivation of the child: overachiever–underachiever; and (*c*) sex of the subject: female–male.

Each subject was seated at a table in the experimental room and was asked to read a set of written instructions (see Brandt, 1971). Included in the instructions was one of the following statements regarding the fictitious student: "His school teacher believes that he is an overachiever since he works very hard [B_1];" or "His school teacher believes that he is an underachiever since he doesn't work very hard [B_2]." The subject was informed that the experiment would be terminated after a given number of trials regardless of the child's progress, and he was instructed to begin the experiment by pressing the "new trial" panel.

Seven blocks of eight trials each were then administered. The fictitious child in the successful condition responded correctly two, three, four, five, six, seven, and seven times, respectively, for each of the seven trial blocks. The unsuccessful child responded correctly two, two, two, three, three, three, and three times, respectively.

After completion of the 56 trials, the subject was asked to complete four evaluation statements by selecting one of the following words or phrases: (*a*) very well, (*b*) well, (*c*) neutrally, (*d*) poorly, or (*e*) very poorly. The evaluation statements were as follows:

1. The child I worked with performed _____.
2. The child I worked with was _____ motivated.

TABLE 1

SIGNIFICANT SOURCES OF VARIANCE FOR THE FOUR DEPENDENT MEASURES

Statement	A	B	BC	ABC
1. Rating of child's performance	$<.01$	$<.01$		$<.01$
2. Rating of child's motivation	$<.01$			$<.01$
3. Subject's rating of enjoyment of task	$<.01$		$<.05$	
4. Subject's rating of own performance	$<.01$			

Note. All values in the table indicate p level. The following were not included in the table because they did not reach significance: C, AB, and AC. Abbreviations: A = performance, B = ascribed motivation, and C = sex of subject.

3. I liked working with this child _____.
4. I feel that as a teacher I performed _____.

Upon termination of the study, 86% of the subjects reported that they believed a child was present and that they were teaching him. The remaining 14% stated that although they were unsure of the existence of the child, they had performed as though a child in fact were present. Analyses revealed no significant differences between these groups; as a result, data from the unsure group were included in the final analyses.

RESULTS

The data of the female education majors did not differ from the data of the female noneducation majors. Therefore, these data were pooled for purposes of statistical analyses.

An analysis of variance of the three independent variables was performed on each of the four response measures (see Table 1). Teachers' attitudes were determined by the level of performance of the fictitious child; the strength of this phenomenon is demonstrated by the fact that all four measures yielded similar results.

Ascribing a motivation level to the child had a significant effect on the teachers' ratings of child performance but not on the other three measures. Inspection of Table 2, which presents cell means for each of the four dependent measures, reveals that subjects judged the child's performance most favorably when the two sets of information (i.e., successful level of performance and overachiever label) were complementary and reflected favorably on the child. They judged his performance least favorably when the two sets of information were complementary and reflected unfavorably on the child. However, when the sources of information were conflicting, such that one was favorable and the other unfavorable, subjects rated the child's performance at intermediate values, indicating the importance of the motivation ascription in determining the teachers' judgments. Similar trends, although not statistically significant, occurred in the other three response measures. Thus, in conflict situations, the ratings indicated that the actual performance of the child carried more weight than the motivation label ascribed to the child.

Figure 1 reveals the differences in mean ratings on the child's performance scale as a function of sex of the teacher. In this triple interaction, both male and female subjects rated successful children more favorably than unsuccessful ones, regard-

TABLE 2

A_iB_j CELL MEANS AND STANDARD DEVIATIONS FOR EACH OF THE FOUR EVALUATION MEASURES

Cell	1		2		3		4	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
A_1B_1	1.50	.50	1.83	.50	1.50	.50	2.42	.39
A_1B_2	2.33	.71	2.25	.71	1.59	.71	2.51	.92
A_2B_1	3.42	.63	2.92	.76	2.75	.84	3.50	.50
A_2B_2	3.75	.59	3.08	1.11	2.50	.85	3.58	1.10

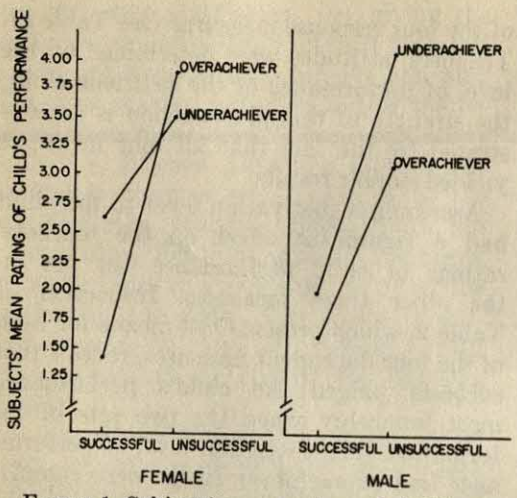


FIGURE 1. Subjects' mean rating of child's performance for levels of apparent motivation at each level of successfulness for female and male subjects.

less of the child's ascribed motivation level. However, male and female subjects differed in that females rated successful overachievers most favorably, unsuccessful overachievers least favorably, and underachievers at intermediate values, while males rated both successful overachievers and successful underachievers favorably and unsuccessful underachievers least favorably.

In evaluating the child's motivational level, the respective ratings of male and female subjects were surprisingly similar to their ratings of the child's performance. It can again be seen in the triple interaction shown in Figure 2 that while successful children were rated more favorably than unsuccessful children by both male and female subjects, female subjects again rated overachievers at the extreme ends of the scale, whereas male subjects again rated both successful overachievers and successful underachievers most favorably and unsuccessful underachievers most unfavorably.

The analysis of variance also indicated a significant Sex of Subject × Motivation Ascription interaction on the subject's rating of his enjoyment of the task. Females rated enjoyment of the task more favorably when working with overachievers than with underachievers. Males enjoyed working with underachievers more than with overachievers. For females, these data are consistent with those presented in Figures 1 and 2 in that females gave more positive ratings to overachievers than to underachievers. Unlike females, males rated overachievers more favorably on the performance and motivation scales, as indicated in Figures 1 and 2, but less favorably on the enjoyment scale. It should be considered that in analyzing four response measures, some effects judged significant could very likely be spurious.

DISCUSSION

This study investigated the effect of ascribed motivation labels and actual performance of fictitious children on teachers' perception of the learning situation and particularly the interaction of these two variables on teachers' attitudes. A second purpose was to investigate whether differences in attitudes existed as a function of the subject's sex.

Ascribed Motivation and Actual Performance of Simulated Child

Performance of the child was found to be the predominant factor in determining the teachers' attitudes, but ascriptions modified the performance effect. This demonstration of the influence of ascriptions on teachers' evaluations is consistent with the results of Medinnis and Unruh (1971) and Rothbart, Dalton, and Barrett (1971), who found that

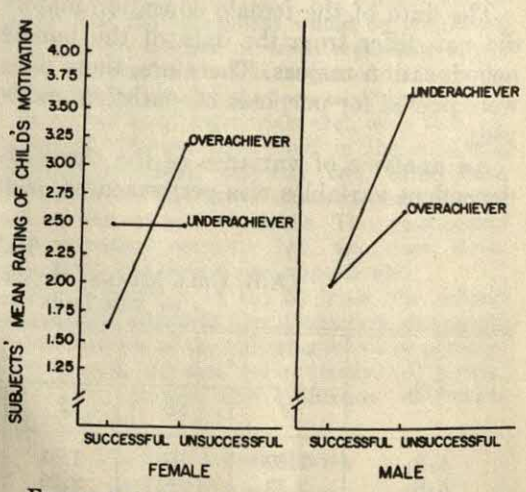


FIGURE 2. Subjects' mean rating of child's motivation for levels of apparent motivation at each level of successfulness for female and male subjects.

teacher performance varied as a function of student ascriptions. However, the present study differs from the latter ones in that personal interaction between teacher and child was prevented so that cues such as physical appearance, racial and socioeconomic status, and behavior characteristics were eliminated. This method also precluded any self-fulfilling-prophecy effect because the teacher had no influence whatsoever on student behavior, which was simulated by the experimenter. The procedure in the present study allowed an assessment of whether the teacher's attitudes were more influenced by her expectancy or by the student's performance, unconfounded by any behavioral interaction.

Predominance of student performance as a determinant of teacher attitudes suggests performance could eventually modify or overcome ascriptions if teachers were not a causative factor in child behavior. Unfortunately, ascriptions do cause differential teacher behavior toward children such that the child's performance is modified and, thus, does not tend to counteract the ascription. In such cases, a self-fulfilling-prophecy situation occurs. Caution should, therefore, be exercised in attaching labels such as "mentally retarded," "slow," or "dull" to a child.

The mixed results obtained in attempts to replicate the *Pygmalion* study reviewed by Brophy and Good (1972) might be partially explained by the results of the present study. Some teachers may be more efficient in controlling their behavior toward students in such a way as to minimize or eliminate the influence of ascriptions. These teachers would then have unconfounded performance to evaluate and thus would be more likely to discover the true performance level of the child. Such a teacher could decrease or prevent an ascription effect. Future studies to evaluate individual teacher efficiency in this area appear warranted.

Sex of Subject

Some differences in attitudes were found between male and female subjects. Although both males and females rated successful children more favorably than unsuccessful ones, females differentiated most in their ratings between successful overachievers and unsuccessful overachievers, whereas males differentiated most in their ratings between successful overachievers and unsuccessful underachievers. There were also sex differences in the types of students the subjects preferred to teach. Males generally enjoyed teaching underachievers, while females generally preferred teaching overachievers.

These findings of sex differences are not surprising in that many similar cases are cited in the literature in which sex of the participants is an important variable in professional–client relationships, such as those between teacher and student (Cosper, 1970), counselor and client (Gamsky & Farwell, 1968; Hebert, 1968), or experimenter and subject (Rosenthal et al., 1964). However, the present study suggests that something more basic than sex role interactions are responsible for these variations. It appears likely that basic differences between male and female preferences or attitudes exist prior to the teacher–student interaction and that these differences influence teacher–student interactions. Good, Sikes, and Brophy (1973) found differences in the behaviors of male and female teachers while interacting with their students. These attitudinal and behavioral findings, after having become reliably established, should have important implications for educational programs.

REFERENCES

Baker, J. P., & Crist, J. L. Teacher expectancies: A review of the literature. In J. D. Elashoff & R. E. Snow (Eds.), *Pygmalion Reconsidered*. Worthington, Ohio: Charles A. Jones, 1971.

Beez, W. Influence of biased psychological reports on teacher behavior and pupil performance. *Proceedings of the 76th Annual Convention of the American Psychological Association*, 1968, **3**, 605–606. (Summary)

Brandt, L. J. Differential response patterns of adults in a teaching situation as a function of motivation and performance levels of children. Unpublished doctoral dissertation, University of Arizona, 1971.

Brophy, J., & Good, T. Teacher expectations: Beyond the *Pygmalion* controversy. *Phi Delta Kappan*, 1972, **54**, 276–278.

Cosper, W. B. An analysis of sex differences in teacher-student interaction as manifest in verbal

and nonverbal behavior cues. Unpublished doctoral dissertation, The University of Tennessee, 1970.

GAMSKY, N. R., & FARWELL, G. F. The effect of client demeanor upon the verbal responses of school counselors. In L. Litwork, R. Getson, & G. Sallzman (Eds.), *Research in counseling*. Itasca, Ill.: G. E. Peacock, 1968.

GOOD, T. L., SIKES, J. N., & BROPHY, J. E. Effects of teacher sex and student sex on classroom interaction. *Journal of Educational Psychology*, 1973, **65**, 74–87.

HEBERT, D. J. Client-counselor personality need and sex similarity. In L. Litwork, R. Getson, & G. Sallzman (Eds.), *Research in Counseling*. Itasca, Ill.: G. E. Peacock, 1968.

MEDINNIS, G., & UNRUH, R. Teacher expectations and verbal communication. Paper presented at the annual meeting of the Western Psychological Association, San Francisco, April 1971.

ROSENTHAL, R., & JACOBSON, L. *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York: Holt, Rinehart & Winston, 1968.

ROSENTHAL, R. et al. Emphasis on experimental procedure, sex of subjects and the biasing effects of experimental hypotheses. *Journal of Projective Techniques and Personality Assessment*, 1964, **28**, 470–473.

ROTHBART, M., DALTON, S., & BARRETT, R. Effects of teacher expectancy on student–teacher interaction. *Journal of Educational Psychology*, 1971, **62**, 47–54.

(Received March 12, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 315-318

CHANGES IN MEMORY STRUCTURE AND RETRIEVAL OVER THE COURSE OF INSTRUCTION[1]

ELIZABETH F. LOFTUS[2] AND GEOFFREY R. LOFTUS

University of Washington

Early research on retrieval of semantic information has provided a reasonably accurate description of retrieval of certain kinds of well-learned material. In the present study, 30 graduate students were asked to produce a type of semantic information; they named psychologists who satisfied certain restrictions. Not only was the speed of naming a psychologist influenced by the order in which restrictions were given, but the effect of order differed for advanced and beginning students. Advanced-student retrieval resembled the pattern observed for well-learned semantic material, while beginning-student retrieval did not. Retrieval was, thus, subtly related to how much instruction a student had completed. These data have implications for the use of reaction time to assess progress in the acquisition of new material.

One of the most fundamental problems confronting today's cognitive psychologist is how to (*a*) represent the knowledge that a person has and (*b*) determine the mechanisms by which a person uses this knowledge. This article addresses the question of how knowledge structures and retrieval mechanisms change during the course of learning new material.

In several recent experiments (Freedman & Loftus, 1971; Loftus, in press; Loftus & Suppes, 1972), a subject was shown a stimulus consisting of a noun category paired with either a letter or an adjective, and his job was to provide a word that satisfied these imposed restrictions. For example, a subject who was presented with the pair *animal–m* might have said mouse, moose, or monkey, among other possibilities. A correct response would have been any word beginning with *m* that named a kind of animal. For *fruit–yellow*, a correct response would have been any *fruit* to which the adjective *yellow* was applicable (e.g., banana, lemon, etc.). A model of semantic memory that accounts for the reaction time data in these experiments assumes that the memory store consists of a large number of interconnected and cross-referenced associative and category networks. According to the model, memory is organized into a complex network composed of categories (e.g., animals) with subsets of each (e.g., birds, dogs) and supersets (e.g., living things). Within each category a variety of subsets exist: Some of them are clusters of items that are highly associated because they have qualities in common (e.g., small animals). Retrieval from this hierarchical structure is assumed to consist of at least two major steps: (*a*) entering the appropriate category and (*b*) finding an appropriate member of that category.

In a newer version of this paradigm, the stimuli have become more complex; for example, in one experiment, subjects were presented with stimuli consisting of a noun category plus both an adjective and a letter (e.g., *animal–small–m*) and had to produce a member of the category that satisfied the two restrictions. That is to say, the response had to be a member of the category that began with the given letter and to which the adjective was applicable (e.g., mouse). Sub-

[1] This research was partially supported by a grant from the National Institutes of Health.

The authors are grateful to William Cole for his assistance in running subjects.

[2] Requests for reprints should be sent to Elizabeth F. Loftus, Department of Psychology, University of Washington, Seattle, Washington 98195.

jects were given the category first, but the order in which the adjective and letter restrictors were presented was varied. On half of the trials, subjects saw the adjective one-half second before the letter (e.g., *animal–small–m*), while on the remaining trials they saw the letter first (e.g., *animal–m–small*). Reaction time was measured from the onset of the last restrictor. The results indicated that a large advantage in reaction time exists when the adjective is presented before the letter. In other words, subjects are considerably faster at naming, for example, an *animal–small–m* than at naming an *animal–m–small*. A discussion of this finding in terms of the network model of semantic memory is as follows. When a subject must produce a category member that satisfies both an adjective and a letter restrictor (e.g., *animal–small–m*), he apparently first enters the category (animals), then restricts himself to the adjective-defined subclass (small animals), and finally he searches there for an item whose name begins with the particular letter requested (*m*). Thus, when the adjective is presented before the letter, the subject can begin the second step earlier.

Taken together, these experiments give us a reasonably good picture of retrieval from extremely well-learned categories. A question of interest is: What happens when categories are not so well learned or are in the process of being learned? One way to study the retrieval of information that is in the process of being learned is to find a situation in which natural learning of categories is taking place. Such a situation exists in all graduate schools of psychology where graduate students are learning, among other things, the names of psychologists. At an institution where one of the authors was teaching, students learn that there are roughly six areas of psychology (learning, perception, memory, personality, social, and developmental) and that various psychologists may be associated with one or more of these areas. Learning to associate psychologists with particular areas of research is tantamount to learning to categorize psychologists with respect to these areas. Different degrees of learning should be evident in people who are in different stages of graduate school.

For the moment let us assume that the number of credits a graduate student has completed is a rough index of the amount of psychology he knows, or the extent to which he has organized psychologists in semantic memory. The major question to which the present research is directed is: Does this objective measure of learning about psychology (number of credits) correlate with the extent to which retrieval (of psychologists) mirrors retrieval of well-learned information such as animals, fruits, etc.

How do we know when retrieval of psychologists mirrors retrieval of well-learned information? The present experiment allowed such a test. Graduate students were asked to produce the names of psychologists. On any given trial, the psychologist named had to satisfy two restrictions: Both (*a*) an area of psychology and (*b*) a letter were shown (e.g., *learning–b*), and the subject had to produce the name of a psychologist that began with the given letter and who was associated in the subject's mind with the given area. For example, a subject who was presented with the stimulus *learning–b* might say Bower, Bourne, or Blodgett, among other possibilities. On half of the trials, subjects saw the area first (e.g., *learning–b*), while on the remaining trials they saw the letter first (e.g., *b–learning*). Reaction time was taken from the onset of the last restrictor. As may be clear without further mention, this area-plus-letter experiment was extremely similar to the adjective-plus-letter experiment in which subjects had to name, for example, an animal that was small beginning with *m*. In the latter experiment, reaction time was much faster when the adjective preceded the letter. If memory for psychologists is as well organized, one might expect the same advantage in reaction time to obtain when the area precedes the letter. If memory for psychologists is not so well organized, no such advantage in reaction time would be expected.

METHOD

Subjects

The subjects were 24 students at the New School for Social Research, New York, New York. Each subject took part in one experimental session that lasted about 30 minutes.

Materials

Six areas of psychology were selected: learning, memory, perception, social, developmental, and personality. Each area was paired with eight different letters, creating 48 unique stimuli. Each stimulus was presented with the area shown first or second (e.g., *learning–b, b–learning*) and with an interval of .5 second between the area and letter. Stimuli were printed on 5 × 8 inch cards.

Each subject received a random permutation of 48 stimuli with the following restrictions: (*a*) A given stimulus (such as *learning–b*) occurred equally often in the area–letter and letter–area conditions; (*b*) half of the stimuli presented to any one subject were in the area–letter order, while the other half were in the opposite order.

Procedure

Each subject was told that the study concerned memory for psychologists and that he was to produce the name of one psychologist on each trial. He was told that he would see items consisting of an area of psychology and a letter and that he should respond with the name of a psychologist that began with the given letter and who was associated in his mind with the given area. He was given examples and told to respond as quickly as possible but to avoid errors.

The subject sat in front of a screen in which was a window covered by half-silvered glass. The index card containing the stimulus was placed in a dark enclosure behind the mirror and was presented by illuminating the enclosure. A microphone was placed in front of the subject, and he responded by speaking into it.

A trial consisted of the following. As a card with the item printed in large type was placed in the darkened enclosure behind the half-silvered mirror, the experimenter said, "Ready" and pressed a button that illuminated the first half of the stimulus. After a .5-second interval, the second half of the stimulus was automatically illuminated, and simultaneously an electric timer with a dc clutch was started. The subject's verbal response activated a voice key that stopped the clock and terminated the trial. A warm-up period of 15 trials preceded the experimental trials.

RESULTS

Only correct responses (56%) to each of the 48 stimuli are included in the following analyses. The most important and interesting result obtained when we separated the subjects according to the number of graduate course credits they had completed. We operationally defined "beginning students" as those who had completed fewer than 40 credits (mean number of credits for these 12 subjects was 27) and "advanced students" as those who had completed more than 45 credits (mean number of credits for these 12 subjects was 54). Two median latencies were obtained for each subject's responses in the two conditions (area–letter and letter–area). For these two conditions, group mean latencies were obtained by averaging the medians separately for the advanced and the beginning students. The results of this analysis are shown in Figure 1. A two-way analysis of variance was performed on the median reaction times in terms of (*a*) order of presentation of the area and letter, and (*b*) type of student. Concerning the main effects: Order of presentation was not significant ($F < 1$), but advanced students responded more quickly than beginning students ($F = 5.60$, $df = 1/22$, $p < .05$). The interaction between these factors was highly significant ($F =$

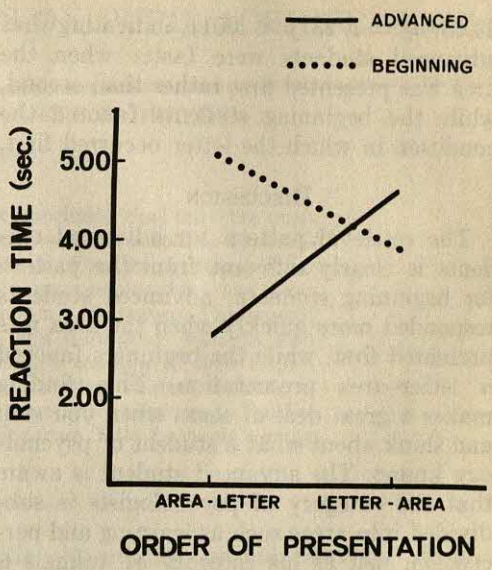


FIGURE 1. Reaction time as a function of order of presentation of area and letter for beginning and advanced students.

33.63, $df = 1/23, p < .001$), indicating that advanced students were faster when the area was presented first rather than second, while the beginning students favored the condition in which the letter occurred first.

Discussion

The retrieval pattern for advanced students is clearly different from the pattern for beginning students; advanced students responded more quickly when the area was presented first, while the beginners favored a letter–area presentation. This finding makes a great deal of sense when you stop and think about what a student of psychology knows. The advanced student is aware that the category of psychologists is subdivided into areas such as learning and perception, just as his category of animals is subdivided. When given the area before the letter, this student can restrict himself to the area-defined subclass and then search for a psychologist whose name begins with the particular letter requested. The beginning student, however, does not have psychology so well organized; the field is not yet neatly subdivided. This student knows a few important names; he probably knows Freud, Skinner, Piaget, and possibly some others. When the letter is given before the area, the beginning student probably begins scanning his list for a name beginning with the particular letter requested and then produces that name almost irrespective of which area is presented. If the student stretches it, Piaget and Skinner could both fit into quite a few different areas of psychology.

Regardless of the exact storage or the exact retrieval mechanisms that these two types of students are using, it is clear that the retrieval patterns observed in the present study are related to how much instruction a person has completed. Furthermore, the retrieval pattern for the advanced students resembles the pattern observed for well-learned material, that is, the more experience a student has had with the field of psychology, the more his retrieval of this information seems to mirror the retrieval of material that we know is well organized and learned. The implication here is that one of the consequences of instruction may be to change a student's retrieval pattern, such that it is more efficient, resembling the retrieval of well-learned material. It appears that we can use reaction time to assess the real impact of instruction in much more subtle ways than we now do. Instruction does more than teach content. In addition, as a person learns new material, his cognitive structure is organized and modified in some way. Reaction time measures such as the ones used in this study can give information about progress being made and ultimately about the process of acquiring new material.

REFERENCES

Freedman, J. L., & Loftus, E. F. The retrieval of words from long-term memory. *Journal of Verbal Learning and Verbal Behavior,* 1971, **10,** 107–115.

Loftus, E. F. Activation of semantic memory. *American Journal of Psychology,* in press.

Loftus, E. F., & Suppes, P. Structural variables that determine the speed of retrieving words from long-term memory. *Journal of Verbal Learning and Verbal Behavior,* 1972, **11,** 770–777.

(Received April 6, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 319-324

MEASURING TEACHER EFFECTS ON PUPIL ACHIEVEMENT[1]

DONALD J. VELDMAN AND JERE E. BROPHY[2]

University of Texas at Austin

Several recent books and articles have concluded that only the quality of the student body, and not the quality of the school or its teaching staff, "makes a difference" on measures of student learning. These studies, however, have measured only presage variables and have used schools rather than teachers as the unit of analysis. The present study, using a sample of 115 second- and third-grade teachers with five or more consecutive years of experience teaching at their respective grade levels, showed that teachers do affect student learning to a degree that is both statistically and practically significant. Teacher effects were especially robust in the data from Title I schools serving disadvantaged populations.

Of late, several books and articles have appeared which argue that "schools don't make a difference." These statements have been based on the Coleman Report (Coleman, et al., 1966) and other investigations which shared several common characteristics that precluded drawing any such inferences. First, they have used *schools* rather than *teachers* as the unit of analysis. Schools are not appropriate units for analysis because they are staffed by teachers of varying ability, and lumping together the data from these individual teachers masks rather than reveals the effects of the quality of schooling. Only data based on the teacher as the unit of analysis can show that some teachers are better than others. Second, to the extent that teaching staffs do differ in quality, schools serving advantaged groups are likely to have better staffs than schools serving disadvantaged groups (Mood, 1970). This factor accentuated the probability of finding that "schools don't make a difference," since better students get better teaching, thus increasing the achievement gap between such students and less gifted students. Third, and most important, one cannot draw conclusions about schooling without measuring it, but this is being done nevertheless. Coleman and his colleagues measured presage variables like years of teaching experience and highest degree obtained, but they did not obtain either process data on teachers' classroom performance or product data showing the student learning gains that individual teachers produced.

The latter technique, employed in this study, is best suited for demonstrating that teachers do in fact make a difference (or, more specifically, that teachers differ in their relative impact on student learning). The usefulness of such data, although seemingly obvious on a commonsense basis, has not been recognized or stressed until recently (Mood, 1970). Furthermore, its validity was seriously challenged by Rosenshine's (1970) review of stability in teacher effectiveness. Rosenshine could locate only five studies containing information on teacher stability over long periods (one semester or more). Of these, one involved

[1] This article is an expanded version of a paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973. The research described was supported by National Institute of Education Contract OE 6-10-108, The Research and Development Center for Teacher Education. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by that office should be inferred. For their assistance in preparing the data and manuscript, the authors wish to thank Marilyn Arnold, Carolyn Evertson, Susan Florence, Kathy Paredes, Kathleen Senior, Jane Sheffield, and John Sheffield.

[2] Requests for reprints should be sent to Jere E. Brophy, Department of Educational Psychology, University of Texas, Austin, Texas 78712.

recruits in an armed service training school and two others came from a study in which a new curriculum was being introduced; therefore, the teachers were not teaching in their accustomed ways. Thus, only two studies reflected teaching by typical teachers working under normal conditions. One study gave no stability coefficient but stated that stability was very low, while the other reported a coefficient of .09.

These data seemed to support the idea that teachers are not stable or consistent in the relative student learning gains they produce and that teacher effectiveness (by this definition, at least) is not a "trait" or stable quality. However, the data from these studies were from unselected teacher samples that may have included substantial proportions of new teachers and/or teachers recently shifted into a new grade. Such teachers are known to be unstable in their teaching behavior (appropriately so, since they are meeting and adjusting to new demands); therefore, they are unlikely to show much stability.

The present study investigated teacher stability in producing student learning gains in a sample of teachers with five or more years of experience teaching at the same grade level. Although few would argue that the amount of gain pupils show on standard achievement tests is the only or even the best measure of teaching effectiveness, it is being used increasingly for this purpose. The present paper concerns methodological considerations involved in obtaining unbiased estimates of teacher influence on pupil achievement and provides data related to the substantive question of whether or not (or how much) teachers "make a difference."

Samples and Measures

All second- and third-grade teachers (N = about 275) in a large Southwestern urban school system were considered for inclusion in a comprehensive investigation of teacher effectiveness, classroom behavior, and personal characteristics. All teachers were female. Teachers selected from the full sample for inclusion in the present study were those who (*a*) had at least five years of teaching experience at their grade, (*b*) had taught the same grade level during the three focal years (1967–1969), and (*c*) had at least 14 children with available data for each of these years (data on 20–30 pupils were available for most classes). The teachers represented 15 Title I (poverty area) and 35 non-Title I schools. The four samples resulting from this selection were as follows:

21 Grade 2, Title I teachers (1,210 pupils);
35 Grade 2, non-Title I teachers (2,168 pupils);
20 Grade 3, Title I teachers (1,216 pupils); and
39 Grade 3, non-Title I teachers (2,744 pupils).

Pupil records were retrieved from school files for each of four successive years of regular fall achievement testing. Grade equivalent scores were obtained for the Metropolitan Achievement Test (MAT) subscales. Different forms of the MAT battery were used with each of the four samples, necessitating separate statistical analyses.

Influences on Predictive Efficiency

It is now generally accepted that residual gain scores are superior to simple pretest–posttest difference scores as measures of teacher influence. What is not clear, however, is the importance of residualizing with more than the simple pretest variable (Cronbach & Furby, 1970). The following series of analyses were designed to explore this problem.

A series of regression models were compared, using (*a*) pretest, (*b*) squared pretest, (*c*) pupil sex, (*d*) year of testing, and (*e*) teacher[3] as predictors of posttest performance. In each comparison, one of these influences was omitted to determine its contribution to prediction of the criterion. Tables 1–4 contain the results of these comparisons, expressed as percentages of criterion variance associated with each influence.

[3] Teachers were represented in the equations by a set of binary variables, one per teacher. The entire set was omitted to estimate teacher influence.

TABLE 1
INFLUENCES ON PREDICTION OF STUDENT LEARNING GAIN MEASURES IN THE GRADE 2, TITLE I SAMPLE

MAT subtest	*M*		Influence (%)				
	Pretest	Posttest	Pretest	Squared pretest	Sex	Year	Teacher
Word Knowledge	1.74	2.41	32.65	.14	.23	.20	6.14
Word Discrimination	1.83	2.81	42.43	.14	.30	1.71	3.55
Reading	1.77	2.52	22.34	.07	1.43	.29	6.90
Arithmetic Computation	1.87	2.79	27.80	.57	.00	1.04	4.50
Verbal average	1.78	2.58	46.77	.13	.47	1.10	5.30
Total	1.83	2.68	42.85	.09	.34	1.56	5.75

Note. Data on the Arithmetic Reasoning subtest were not available for this sample. Abbreviation: MAT = Metropolitan Achievement Test.

These data suggested the following conclusions:

1. Inclusion of a squared score variable to permit curvilinear regression added little to the precision of the estimates, although it was slightly more influential in Grade 3 than in Grade 2.

2. Inclusion of pupil sex in the model added very little to predictive efficiency, even though girls significantly outperformed boys at these grade levels. The sex difference was included in the prescores, however, so that its influence was negligible with prescores controlled.

3. Systematic differences among the three years of testing were trivial. This was expected, since there was no known reason to believe that any yearly differences in residual gain would appear.

4. Inclusion of the teacher variable usually yielded a significant, and often a substantial, increase in predictive efficiency. In short, teachers did make a difference, although pupil prescores were usually the strongest predictors by a considerable margin.

5. The influences of sex, year, and teacher appeared to be stronger in Title I than in non-Title I schools. Although the reasons for the year-effect differences were unclear, the sex- and teacher-effect differences were readily interpretable. Sex differences, including sex differences in school achievement, are more extreme in lower- than in higher-socioeconomic-status populations (Hess, 1970), so that a greater sex effect was expected in the Title I schools. Similarly, the differential teacher effect was expected on the basis of what is known about the relationship of ability and achievement

TABLE 2
INFLUENCES ON PREDICTION OF STUDENT LEARNING GAIN MEASURES IN THE GRADE 2, NON-TITLE I SAMPLE

MAT subtest	*M*		Influence (%)				
	Pretest	Posttest	Pretest	Squared pretest	Sex	Year	Teacher
Word Knowledge	2.50	3.71	61.01	.30	.03	.03	2.69
Word Discrimination	2.82	3.70	58.11	.41	.31	.04	2.95
Reading	2.51	3.62	53.17	.01	.01	.04	2.97
Arithmetic Computation	2.52	3.18	29.93	.92	.16	.22	5.66
Arithmetic Reasoning	2.52	3.25	40.12	.00	.00	.07	4.22
Verbal average	2.61	3.67	71.72	.24	.00	.03	2.28
Quantitative average	2.52	3.22	42.30	.24	.03	.05	5.02
Total	2.57	3.44	70.38	.00	.02	.00	3.04

Note. Abbreviation: MAT = Metropolitan Achievement Test.

TABLE 3

INFLUENCES ON PREDICTION OF STUDENT LEARNING GAIN MEASURES IN THE GRADE 3, TITLE I SAMPLE

MAT subtest	*M*		Influence (%)				
	Pretest	Posttest	Pretest	Squared pretest	Sex	Year	Teacher
Word Knowledge	2.51	3.27	37.15	.90	.10	.33	17.64
Word Discrimination	2.82	3.34	44.33	.35	.65	1.85	4.96
Reading	2.62	3.25	34.36	.36	.21	.16	10.86
Arithmetic Computation	2.76	3.44	29.89	.03	.41	1.18	9.34
Arithmetic Reasoning	2.76	3.05	26.63	.86	.36	.36	4.04
Verbal average	2.65	3.29	51.28	1.00	.16	.94	12.44
Quantitative average	2.76	3.25	34.40	.14	.47	.87	6.31
Total	2.70	3.27	50.08	1.16	.41	1.44	9.83

Note. Abbreviation: MAT = Metropolitan Achievement Test.

differences to socioeconomic status. At a given age level, the cognitive abilities and school achievement of lower-socioeconomic-status children are less advanced than those of their higher-socioeconomic-status peers. This can be seen in the present data by comparing (*a*) the means of Table 1 versus those of Table 2 with (*b*) the means of Table 3 versus those of Table 4. Among other things, this means that the more advantaged children were better able to learn on their own and/or from one another and thus were less dependent upon the teacher for their degree of success in mastering the curriculum—hence, the greater teacher effect in Title I schools.

6. Teacher impact appeared to be stronger on verbal skills than on quantitative skills in Title I schools and vice versa in non-Title I schools. This was probably an elaboration of the factor described in Paragraph 5; socioeconomic-status differences are greatest on measures of verbal skills.

7. Teacher impact was stronger in Grade 3 than in Grade 2 in Title I schools but about equal in non-Title I schools. Again, this was probably an elaboration of the factor described in Paragraph 5; by Grade 3, the socioeconomic-group differences, and thus the opportunity for teacher impact in Title I schools, were larger than they were in Grade 2 (note the larger differences in Grade 3 postscores than in Grade 2 postscores in Tables 2 and 4).

8. Predictability of posttest scores of pupils was generally greater in Grade 3 than in Grade 2, in non-Title I than in Title I schools, and on verbal than on

TABLE 4

INFLUENCES ON PREDICTION OF STUDENT LEARNING GAIN MEASURES IN THE GRADE 3, NON-TITLE I SAMPLE

MAT subtest	*M*		Influence (%)				
	Pretest	Posttest	Pretest	Squared pretest	Sex	Year	Teacher
Word Knowledge	3.67	4.85	64.98	.67	.00	.02	2.16
Word Discrimination	3.66	4.62	67.08	.99	.03	.09	1.29
Reading	3.52	4.65	57.09	1.47	.00	.15	1.41
Arithmetic Computation	3.14	4.13	33.84	1.03	.07	.07	6.94
Arithmetic Reasoning	3.23	4.23	50.55	.00	.02	.12	4.28
Verbal average	3.62	4.71	76.04	1.16	.00	.08	1.08
Quantitative average	3.19	4.18	54.89	.14	.00	.12	5.23
Total	3.40	4.44	75.92	.39	.00	.12	2.16

Note. Abbreviation: MAT = Metropolitan Achievement Test.

quantitative measures. The first and third conclusions have been discussed above. The greater predictability in non-Title I schools resulted from the higher correlations between pre- and postscores in these schools. This, in turn, was due most likely to the combined effects of (*a*) the great variance in scores in non-Title I schools, which made for higher correlations when coefficients were not corrected for attenuation and (*b*) the more advanced verbal and math skills of children in the non-Title I schools, which enabled them to answer correctly more often and to guess less often than children in Title I schools, whose scores very likely contained more error variance due to guessing or other response sets.

Consistency of Teacher Impact

The next step of the analysis addressed the question of the degree to which individual teachers' influence on child gain was consistent across three successive years, and hence, across classes of pupils.

Residual gain scores for all pupils were obtained, using only simple pretest scores as covariates. These were then averaged for each teacher for each of her three classes. These average residual gains were then used to compute intraclass correlations among the three years for each of the four samples of teachers. Intraclass correlations (Ebel, 1951) provided indices of the consistency of pupil gain within teachers, across classes of pupils. The results of this analysis are shown in Table 5. With the exception of the second grade, Title I sample, in which none of the coefficients were statistically significant, it was apparent that three-year averages[4] were reasonably reliable estimates of teacher impact on student learning.

TABLE 5
Intraclass Correlations Across Three Years

MAT subtest	Grade 2		Grade 3	
	Title I	Non-Title I	Title I	Non-Title I
Word Knowledge	.43	.66*	.78*	.63*
Word Discrimination	.36	.74*	.26	.49*
Reading	.24	.66*	.50*	.23
Arithmetic Computation	.00	.48*	.63*	.80*
Arithmetic Reasoning	—	.61*	.27	.64*
Verbal average	.35	.71*	.65*	.38*
Quantitative average	—	.59*	.50*	.75*
Total	.19	.69*	.54*	.65*

Note. Abbreviation: MAT = Metropolitan Achievement Test.
* $p < .05$.

Conclusions

The data show that reasonably stable estimates of teacher influence can be obtained from standardized achievement measures of pupil performance when sample selection procedures eliminate new teachers and teachers who have recently switched grades. The increasing use of the team approach in elementary schools, however, lessens the practical interest of such measures. Also, although the stability coefficients from this study were considerably higher than those located by Rosenshine (1970), they were not high enough to justify the use of residual gains on such measures for teacher accountability purposes (Brophy, 1973).

The differences between Title I and non-Title I schools were consistent with the theoretical position that the school is relatively more important, compared to the home, in determining the achievement levels of economically disadvantaged pupils than it is for advantaged pupils. This, in turn, suggested that the quality of teaching is more crucial in such settings than in advantaged schools.

A subsample of these teachers, selected because they showed the greatest consistency across four years in the degree of student learning gains they produced, is presently being studied in an effort to establish those personal traits and classroom behaviors which are associated with teacher effectiveness in producing learning gains (Brophy & Evertson, 1973; Evertson & Brophy, 1973; Peck & Veldman, 1973). Having established that teachers have differential impacts in determining student scores on

[4] These intraclass coefficients (Ebel, 1951) concern three-year averages and are *not* averages of the three possible two-year correlations.

product measures, we are now attempting to identify the presage and process variables associated with these differences.

REFERENCES

BROPHY, J. Stability in teacher effectiveness. *American Educational Research Journal,* 1973, **10,** 245–252.

BROPHY, J., & EVERTSON, C. Low inference observational coding measures and teacher effectiveness. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

COLEMAN, J., et al. *Equality of Educational Opportunity.* Washington, D.C.: U.S. Office of Health, Education, and Welfare, 1966.

CRONBACH, L. J., & FURBY, L. How we should measure "change"—Or should we? *Psychological Bulletin,* 1970, **74,** 68–80.

EBEL, R. Estimation of the reliability of ratings. *Psychometrika,* 1951, **16,** 407–423.

EVERTSON, C., & BROPHY, J. High inference behavioral ratings as correlates of effective teaching. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

HESS, R. Class and ethnic influences upon socialization. In P. Mussen (Ed), *Carmichael's manual of child psychology.* (3rd ed.) Vol. 2. New York: Wiley, 1970.

MOOD, A. Do teachers make a difference? In, *Do Teachers Make a Difference? A Report on Recent Research on Pupil Achievement.* (U.S. Office of Education 58042) Washington, D.C.: U.S. Government Printing Office, 1970.

PECK, R., & VELDMAN, D. Personal characteristics associated with effective teaching. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

ROSENSHINE, B. The stability of teacher effects upon student achievement. *Review of Educational Research,* 1970, **40,** 647–662.

(Received April 6, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 325-328

A FOLLOW-UP STUDY OF TEACHER EXPECTANCY EFFECTS[1]

EDWARD J. O'CONNELL, JEROME B. DUSEK,[2] AND RICHARD J. WHEELER

Syracuse University

Further longitudinal data were presented from a study of teacher-bias and teacher-expectancy effects on elementary school childrens' achievement test performance. The analyses of the data, Stanford Achievement Test (SAT) scores from the beginning and middle of the 1972–1973 academic year, revealed that the teacher-bias manipulation had no effect and that the teacher-expectancy manipulation was strongly related to SAT performance. Both findings were consistent with the results previously reported. Since the children were at new grade levels and had new teachers, the data were interpreted as indicating that teachers do not bias the learning of children but are good long-term predictors of children's academic capabilities.

The purpose of this paper is to report further longitudinal data from an investigation of teacher-bias and teacher-expectancy effects on elementary school children's Stanford Achievement Test (SAT) performance. In the earlier report (Dusek & O'Connell, 1973), it was argued that rather than investigate teacher-bias effects by manipulating teachers' expectancies for general intellectual development (e.g., Claiborn, 1969; Rosenthal & Jacobson, 1968), a more meaningful approach would be to manipulate teachers' expectancies in terms of children's performance in specific academic subjects and then to measure children's achievement in these areas. The value of measuring the relationship between teachers' own self-generated expectancies and children's SAT performance was also pointed out in the original report. The rationale for this approach was based on research (e.g., Rist, 1970) indicating teachers form their own expectancies regarding students' potential in terms of performance in academic subject areas.

In order to investigate both teacher-bias and teacher-expectancy effects, Dusek and O'Connell (1973) administered the SATs, disguised as tests aimed at predicting academic potential, to two second- and two fourth-grade classrooms. At the same time, the teachers were asked to rank the children in their room on the basis of expected year-end performance levels for language and arithmetic skills. This ranking was taken as a measure of teacher expectancy for students' performance. The names of 8 of the first 16 students ranked by the teacher in each classroom were given to the teacher along with a statement indicating that this group should show large academic gains in language and arithmetic skills during the coming year. The remaining 8 children formed a control group. This was taken as a manipulation of teacher bias, that is, a measure of expectancy effects based on the intervention of the principal investigator.

The dependent variable was the children's SAT performance at the beginning, middle, and end of the academic year. The findings were very clear. First, there was no evidence of a teacher-bias effect on SAT performance for any of the three testing occasions. Second, teacher-expectancy effects were very strongly related to SAT performance on all three testing occasions. Finally, teacher-bias

[1] The project reported herein was performed pursuant to Grant OEG 2-71-0516 from the U.S. Office of Education, Department of Health, Education, and Welfare. The opinions expressed herein, however, do not necessarily reflect the position or policy of the U.S. Office of Education, and no official endorsement by the U.S. Office of Education should be inferred. The authors are indebted to James McGee, principal, and the second-, third-, fourth-, and fifth-grade teachers of Clinton Elementary School for their kind cooperation.

[2] Requests for reprints should be sent to Jerome B. Dusek, Department of Psychology, Syracuse University, Syracuse, New York 13210.

and teacher-expectancy effects did not interact. Clearly, simply telling teachers that students would perform well was not enough to alter the students' SAT performance; this was a failure to replicate the findings of Rosenthal and Jacobson (1968). Teachers' own judgments, however, were very good predictors of students' performance. Based on the data in the original report, Dusek and O'Connell (1973) argued that these latter teacher-expectancy effects were, in all probability, a reflection of the teachers' ability to accurately estimate the academic potential of the children in their rooms rather than a bias effect in the Rosenthal and Jacobson (1968) sense.

In the present paper, data bearing on the long-term effects of teacher bias and teacher expectancy are presented. At the beginning and middle of the 1972–1973 academic year, the SATs were again administered to the students, who were now in the third- and fifth-grades. The purpose was to determine if teacher-bias or teacher-expectancy effects would be present in SAT performance even though the students had now advanced to new grade levels and had new teachers.

Method

Subjects

Of the original 64 subjects, 38 were still available for testing during the 1972–1973 academic year,[3] including 22 now in the third grade (mean chronological age was 8.4 years) and 16 now in the fifth grade (mean chronological age was 10.0 years). A more complete description of the sample and the basis for its selection may be found in Dusek and O'Connell (1973).

Tests

In September and January of the 1972–1973 academic year, the following subtests from the SAT Primary 2 and Partial Intermediate 1 batteries were administered to the third- and fifth-graders, respectively: Word Reading, Paragraph Meaning, Spelling, Arithmetic Computation, and Arithmetic Concepts. An alternate form of each test battery was administered on each testing occasion. As during the previous year, tests were disguised as measures of academic potential.

Procedure

The procedures were essentially the same as those described by Dusek and O'Connell (1973). The third- and fifth-garde teachers were told that the tests to be administered were developed to predict academic potential in language and arithmetic skills and were being administered for standardization and validation of the data collected the previous year. Since the tests were administered under this guise and since the primary interest was in the long-term effects of teachers' expectancies on performance, these teachers were not asked to rank the children.

Design and Analysis

There were three factors in the design: grade level, experimental condition, and teacher ranking from the 1971–1972 academic year. Since 26 of the original 64 subjects were no longer available for testing, there was not an equal number of subjects in each cell of the design. For this reason, the data were analyzed by the multiple regression technique described by Cohen (1968) and Overall and Spiegel (1969). The application of this technique to the present experiment was spelled out in detail in the initial report.

Results

As in the previous report, the dependent variables were total SAT 4 and SAT 5 scores, from the testing at the beginning and middle of the year, respectively. In addition, the SAT 1 total score was analyzed for the 38 subjects still available from the original sample. The results of the multiple regression analyses are summarized in Table 1. The means associated with the main effects in each analysis are presented in Table 2. The analysis of the SAT 1 scores for the reduced sample revealed effects parallel to the original SAT 1 analysis, that is, the only

[3] In order to examine whether the 13 subjects lost from June 1972 to January 1973 were different from the 38 subjects remaining at the conclusion of the experiment, both the design variables and the criterion variables (SAT 1, 2, and 3 scores) were entered into a multiple discriminant analysis. The F ratio approximation to Wilk's lambda was 1.24 ($df = 10/40$), too small to reject the multivariate null hypothesis of group equivalence on the vector of design and criterion variables. In addition, chi-square analyses indicated no differential loss of subjects as a function of sex of subject, teacher ranking, time of SAT administration, experimental versus control condition, or grade level. These analyses clearly demonstrated no differential loss of subjects as a function of design variables. In addition, the SAT 1, SAT 2, and SAT 3 scores of the 13 subjects lost do not differ from the SAT 1, SAT 2, and SAT 3 scores of the 38 subjects available.

TABLE 1

MULTIPLE CORRELATIONS AND F RATIOS FOR MAIN EFFECTS

Main effect	Stanford Achievement Test		
	1	4	5
Experimental condition			
R^2 full	.3958	.4229	.4598
R^2 reduced	.3958	.4213	.4537
Difference	.0000	.0016	.0061
F[a]	.0000	<1.0000	<1.0000
Grade level			
R^2 full	.3958	.4229	.4598
R^2 reduced	.3953	.3093	.4594
Difference	.0005	.1136	.0004
F[a]	<1.0000	6.69*	<1.0000
Teacher ranking			
R^2 full	.3958	.4229	.4598
R^2 reduced	.0272	.0714	.0105
Difference	.3686	.3516	.4493
F[a]	20.74**	20.71**	28.28**

[a] $df = 1/34$.
* $p < .01$.
** $p < .001$.

significant effect was teacher ranking. Teacher ranking was also a significant effect in the analyses of SAT 4 and SAT 5 scores. In all three cases, students who were ranked higher by the teacher had higher SAT total scores than students who were ranked lower. The correlations between SAT scores and teacher ranking were −.63, −.55, and −.67 for SAT 1, SAT 4, and SAT 5, respectively.

Grade level was a significant effect only in the analysis of SAT 4. The fifth graders had a higher mean score than the third graders. The experimental condition produced no significant effects in any of the analyses. None of the interactions was significant.

DISCUSSION

The results presented above substantiate the findings and interpretations of the previous report in several ways. First, the teacher-bias manipulation had no effect on SAT 4 or SAT 5 performance. Taken together, the findings in the two reports clearly indicate that simply telling teachers that students will perform well is not enough to alter students' performance. These findings, as well as those of other investigators (e.g., Claiborn, 1969), are clear failures to replicate Rosenthal and Jacobson's (1968) research, and they argue that teachers do not bias either the intellectual development or achievement of children of elementary school age.

Second, teacher ranking was strongly and consistently related to the children's SAT 4 and SAT 5 performance, just as it was to their SAT 1, SAT 2, and SAT 3 performance in the initial report. Clearly, the relationship between teacher ranking and SAT performance is not a teacher-bias effect in the Rosenthal and Jacobson (1968) sense. First, as Dusek and O'Connell (1973) noted, it is unlikely that the relationship of teacher ranking to the SAT 1 scores would be as strong after only two weeks of classes if the rankings were, in large part, formed from bases irrelevant to the students' academic ability. Second, during the academic year of SAT 4 and SAT 5 testing, the students were no longer under the tutelage of the teacher who made the rankings. If the rankings were a reflection of teacher bias, the strength of the relationship to SAT performance should have been dissipated. Obviously, this was not the case.

Dusek and O'Connell (1973) argued that the effects associated with teacher ranking were simply a reflection of the teachers' ac-

TABLE 2

MEAN STANFORD ACHIEVEMENT TEST (SAT) SCORES FOR EACH MAIN EFFECT

Main effect	M SAT score[a]		
	1	4	5
Condition			
Experimental	59.16	75.53	84.21
Control	54.26	68.58	81.11
Grade			
Third	58.77	65.09	84.91
Fifth	53.88	81.63	79.56
Teacher ranking[b]			
1–4	78.00	101.75	116.50
5–8	57.88	67.13	86.50
9–12	57.18	76.36	77.82
13–16	39.91	49.73	60.09

[a] SAT 1 = October 1971, SAT 4 = September 1972, and SAT 5 = January 1973.
[b] Teacher ranking was entered as a continuous variable in the multiple regression analysis. The data are grouped here simply for convenience.

curacy at estimating the academic ability levels of the children in their classrooms. The longitudinal data presented above support this hypothesis. What is not known, at present, are the bases teachers use to form expectancies for students' performance. The data in the present paper suggest that teachers' expectancies are based on criteria relevant to academic performance rather than on criteria related primarily to social class, as was reported by Rist (1970). If this is the case, teachers do not bias the education of children.

REFERENCES

CLAIBORN, W. L. Expectancy effects in the classroom: A failure to replicate. *Journal of Educational Psychology*, 1969, **60**, 377–383.

COHEN, J. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 1968, **70**, 426–443.

DUSEK, J. B., & O'CONNELL, E. J. Teacher expectancy effects on the achievement test performance of elementary school children. *Journal of Educational Psychology*, 1973, **65**, 371–377.

OVERALL, J. E., & SPIEGEL, D. K. Concerning least squares analysis of experimental data. *Psychological Bulletin*, 1969, **72**, 311–322.

RIST, R. Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard Educational Review*, 1970, **40**, 411–451.

ROSENTHAL, R., & JACOBSON, L. *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston, 1968.

(Received May 7, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 329-332

BLACK AND WHITE CHILDREN'S COMPREHENSION OF STANDARD AND NONSTANDARD ENGLISH PASSAGES

SAMUEL J. MARWIT[1] AND GAIL NEUMANN

University of Missouri—St. Louis

Two black and two white examiners administered standard English and nonstandard English forms of the Reading Comprehension section of the California Reading Test to 60 black and 53 white second graders. The hypotheses that black subjects comprehend nonstandard English materials better than those in standard English and that white subjects comprehend standard English materials better than those in nonstandard English were not supported. Within each form, white subjects generally obtained higher scores than black subjects, and within each race, standard English presentations generally resulted in higher scores than nonstandard English presentations. Black subjects performed as well as white subjects under the white examiner – standard English condition only. Results were discussed in relation to other studies of the linguistic interference hypothesis.

Much of the current sociolinguistic literature (Williams, 1970) has been concerned with the notion that black children, especially those of low socioeconomic background, speak a "nonstandard" form of English which differs from that used by their white peers. Ethnographical (Bauman, 1971) and empirical (Marwit, Marwit, & Boswell, 1972) investigations have documented the internal consistency of black nonstandard English and the reliability and predictability of its grammatical and phonological rules. The general conclusion has been that black nonstandard English constitutes a different, rather than a deficient, language in relation to standard English and that it needs to be accepted as such.

Recently, attention has focused on the implications of these findings for the education of the black child. Some (Saville, 1971) contend that emotional and cognitive interference phenomena are operating unfairly upon the nonstandard, English-speaking black child, since he is required to perform in a standard, English-oriented school setting and to compete with others who have used standard English all of their lives. While some criticism has focused on the nonacceptance of spoken and written nonstandard English in the classroom (Johnson, 1971), much more has been directed against the nonavailability of "black-language" reading materials for black students. Wolfram (1970) and others (Baratz & Shuy, 1969) have suggested that the low level of reading skills traditionally displayed by black children results from the linguistic interference produced by the standard English format of texts and that reading proficiency, especially comprehension, would be greatly enhanced by translating standard English materials into nonstandard English. Support for the direct relationship between reading comprehension and similarity of written and oral language patterns has been provided by Ruddell (1965). Yet, findings by Nolen (1972) and Torrey (1971) which directly addressed the black child's comprehension of written standard English have refuted this. The present study is designed to investigate this relationship further by having black and white examiners administer standard and nonstandard English forms of a formal reading comprehension test to an interracial population which had previously (Marwit et al., 1972) demonstrated sociolinguistic

[1] Requests for reprints should be sent to Samuel J. Marwit, Department of Psychology, University of Missouri, St. Louis, Missouri 63121.

differences. On the basis of earlier speculations (Marwit et al., 1972), it is hypothesized that black children obtain higher reading comprehension scores when the instructions, test materials, and test questions are presented in a nonstandard English format than when they are presented in a standard English format, that the reverse will be true for white children, and that no subject-race differences will occur among children reading passages written in their native oral dialect.

Method

Subjects

Subjects were 60 black and 53 white second graders from a St. Louis County public school system in which significant subject-race differences in oral language had been previously demonstrated (Marwit et al., 1972). Two white and two black undergraduate males served as examiners, each testing one group of subjects using a standard English format and a second group using a nonstandard English format. Random assignment of subjects within each race to each of the four conditions (black examiner – standard English, black examiner – nonstandard English, white examiner – standard English, white examiner – nonstandard English) resulted in 15 black subjects per condition, 13 white subjects per each of the first three conditions, and 14 white subjects in the last condition.

Test Materials

Test materials consisted of standard English and nonstandard English forms of the Reading Comprehension section of the California Reading Test for first- and second-grade students.[2] The nonstandard English form was prepared by asking two black St. Louis-born teachers to translate the instruments' instructions, reading passages, and test questions into the language of the black, St. Louis school child. Interinterpreter agreement was uniformly high and confirmed expectations based upon earlier research findings (Marwit et al., 1972).

An example of a standard English paragraph is, "Roy plays with the cow. The cow's name is Spotty. Nancy plays with the goat. Its name is Blacky. [p. 7]." The nonstandard translation reads, "Roy, he play with the cow. The cow name Spotty. Nancy, she play with the goat. Its name Blacky."

[2] Adapted from the California Reading Test, Lower Primary, Form X, devised by Ernest W. Tiegs and Willis W. Clark. Copyright © 1957, 1963 by McGraw-Hill, Inc. By permission of the publisher, CTB/McGraw-Hill, Del Monte Research Park, Monterey, California 93940. All Rights Reserved. Printed in U.S.A.

TABLE 1
Mean Reading Comprehension Scores

Subject	Examiner	Format			
		Standard English		Nonstandard English	
		M	*SD*	*M*	*SD*
Black	Black	4.47	2.97	4.60	2.80
	White	8.47	4.26	4.47	2.50
White	Black	9.54	4.10	7.15	4.34
	White	8.31	4.59	8.57	3.99

Results

Table 1 presents the mean reading comprehension scores for black and white subjects tested by black and white examiners using standard and nonstandard English formats. A 2 × 2 × 2 (Subject Race × Examiner Race × Format) analysis of variance resulted in three significant effects: a main effect of subject race ($F = 16.95$, $df = 1/105$, $p < .001$) with white subjects receiving higher comprehension scores than black subjects; a main effect of format ($F = 4.47$, $df = 1/105$, $p < .05$) with subjects achieving higher scores under standard English than under nonstandard English format; and a triple interaction ($F = 5.79$, $df = 1/105$, $p < .05$) produced by black subjects obtaining significantly higher scores under the standard English format when tested by white examiners than when tested by black examiners ($F = 8.62$; $F'_{(.05)} = 8.34$; Sheffé, 1953), while no comparable differences occurred for black subjects under nonstandard English format or for white subjects under either format.

Discussion

The present results failed to support the hypotheses. Black second graders did not comprehend reading materials better when they were presented in a nonstandard English format as opposed to a standard English format nor did their white peers comprehend standard English significantly better than nonstandard English. Furthermore, subject-race differences were generally maintained despite changes in format.

These findings refute the contention of those who hold that linguistic interference between written and spoken language is responsible for most of the reading difficulties encountered by black children (Baratz & Shuy, 1969) and suggest that alternative explanations be explored.

While failing to support the present hypotheses, these results were not unsupported by others in the sociolinguistic and educational literature. They were confirmed by the findings of a number of investigations dealing with the black child's ability to comprehend both written (Nolen, 1972; Torrey, 1971) and spoken (Hall, Turner, & Russell, 1973) standard English. In each of these investigations, the presence of a reliable, nonstandard English dialect as the black child's oral language pattern was documented, as was the failure of this oral dialect to interfere with the comprehension of standard English. Where white subjects were employed, similar nonsignificant Subject Race × Format interactions were obtained. These findings, taken together with those of the present study, suggest that either linguistic discrepancies due to subject race are not pronounced enough to impair comprehension or are simply unimportant to the process of understanding. It has yet to be demonstrated that direct oral–oral or oral–visual correspondence is essential for comprehension of verbal and written messages among speakers whose basic vocabularies and sentence structures are roughly equivalent.

The significant main effects of subject race and of format were not by themselves relevant to the present hypotheses. The means comprising the two-way interaction of these variables, while not significantly different from each other, were interesting. Within each format, white subjects obtained higher comprehension scores than black subjects. This was contrary to the hypotheses and has already been discussed. Within each subject race, higher scores were elicited by the standard English format than by the nonstandard English format, with the discrepancy being more pronounced among black subjects. It is not surprising that white children performed slightly better when exposed to standard English, since this is the language they have heard and spoken all their lives. However, black children's achievement of higher test scores under standard English conditions was unexpected; this might be a function of either their familiarity with standard English as the expected and accepted language of the classrooms and therefore the one which demanded performance or, conversely, a function of their distrust of nonstandard English in a setting where it was rarely, if ever, used and almost never rewarded (Johnson, 1971). Observation of the means (Table 1) entering into the Subject Race × Examiner Race × Format interaction provides further support for either of these interpretations. This interaction was the result of black subjects performing as well as white subjects under the single circumstance of a white authority figure using standard English. If it can be assumed that the results of this condition were not due to chance, then it is reasonable to suggest that black second graders can comprehend written material as well as white second graders but manifest this only under specified conditions. Research designed to investigate the facilitating and suppressing effects of various subject race – examiner race combinations and of various expectancy, confirmation – disconfirmation circumstances is being prepared and will hopefully provide more information.

REFERENCES

Baratz, J., & Shuy, R. W. (Eds.) *Teaching black children to read.* Washington, D. C.: Center for Applied Linguistics, 1969.

Bauman, R. An ethnographic framework for the investigation of communicative behaviors. *Asha,* 1971, **13,** 334–340.

Hall, V. C., Turner, R. R., & Russell, W. Ability of children from four subcultures and two grade levels to imitate and comprehend crucial aspects of standard English: A test of the different language explanation. *Journal of Educational Psychology,* 1973, **64,** 147–158.

Johnson, H. H. Teacher attitude and ghetto language. *Viewpoints,* 1971, **47,** 73–81.

Marwit, S. J., Marwit, K. L., & Boswell, J. J. Negro children's use of nonstandard grammar. *Journal of Educational Psychology,* 1972, **63,** 218–224.

Nolen, P. S. Reading nonstandard dialect ma-

terials: A study at grades two and four. *Child Development*, 1972, **43**, 1092–1097.

RUDDELL, R. B. The effect of oral and written patterns of language structure on reading comprehension. *Reading Teacher*, 1965, **18**, 270–275.

SAVILLE, M. R. Interference phenomena in language teaching: Their nature, extent, and significance in the acquisition of standard English. *Elementary English*, 1971, **48**, 396–405.

SCHEFFÉ, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 1953, **40**, 87–104.

TORREY, J. W. Teaching standard English to speakers of other dialects. In G. E. Perren & J. L. M. Trim (Eds.), *Applications of linguistics: Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969*. Cambridge, England: Cambridge University Press, 1971.

WILLIAMS, F. (Ed.) *Language and poverty: Perspectives on a theme*. Chicago: Markham, 1970.

WOLFRAM, W. A. Sociolinguistic alternatives in teaching reading to nonstandard speakers. *Reading Research Quarterly*, 1970, **6**, 9–33.

(Received May 14, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 333-338

VARIABILITY IN CHILDREN'S COMPREHENSION OF SYNTACTIC STRUCTURES[1]

ALAN M. LESGOLD[2]

University of Pittsburgh

Data are presented that challenge the difficulty ordering for anaphoric syntax (e.g., pronouns) proposed by Bormuth, Manning, Carr, and Pearson. It is suggested that any such difficulty ordering which results from tests of the form proposed by Bormuth has uncontrolled variability due to semantic factors that have yet to be carefully analyzed and controlled.

Bormuth, Manning, Carr, and Pearson (1970) have reported results suggesting that stable difficulty orderings could be obtained for three classes of syntactic forms containing a total of 55 separate variants. There were difficulty differences both between and within the three classes (intrasentence, intersentence, and anaphora[3]), from which they argued that these categories may be related to stages in a learning hierarchy for syntax comprehension skills. Further, there have been recent suggestions that additional experimentation of this sort, combined with linguistic analysis, will result in discovery of such a hierarchy (e.g., Carroll, 1972; Frase, 1972). The present results—a partial replication of their work—suggest that this approach is unlikely to succeed unless such factors as semantics and constraints on information-processing capacity are concurrently considered.

The procedure Bormuth et al. (1970) used in determining the difficulty of the various syntax forms was simple and, at first glance, straightforward. For a given anaphora form, for example, one sentence was written and then a second was constructed with some reference back to the first (e.g., *John went to the store. He bought a pear.*). A paragraph was then constructed around the two-sentence cluster. A question was generated for the paragraph by substituting a *wh-* word (Bormuth, 1970) for the anaphora (e.g., *Who bought a pear?*). Finally (for anaphora only), multiple-choice alternative answers were written. The item thus consisted of a paragraph, a question, and a number of alternative answers from which to choose.

There is a basic problem in using procedures of this sort for measuring the difficulty of one syntactic structure relative to another. This is the confounding of syntax with semantics. Consider a potential comparison between a personal pronoun structure (e.g., *he*) and a pro-clause form (in which *that* or *so* might stand for an entire clause). The two syntactic forms occur in different semantic contexts: there is no semantic (deep) structure that can alternatively be expressed as either a personal pronoun or a pro-clause. Now this does not necessarily mean that a difficulty ordering such as that of Bormuth et al. should be discounted. After all, the confounding of syntax and semantics may, for such purposes, be complete.

If the Bormuth et al. (1970) difficulty ordering is consistently replicable, then we can still look for a learning hierarchy re-

[1] This research was supported by the Learning Research and Development Center, University of Pittsburgh, through funds from the National Institute of Education. The experiment was conducted by Hildrene De Good. Karen Block, Roberta Golinkoff, and Charles Perfetti read an earlier draft and helped in improving it.

[2] Requests for reprints should be sent to Alan M. Lesgold, LRDC, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

[3] *Anaphora* is the term used to denote a structure in a sentence, for example, a pronoun, that derives its meaning from a previous sentence or an earlier part of the present sentence.

lated to that data, even though we would not know, initially, the relative roles of syntax and semantics. On the other hand, it may be that the difficulty ordering is not stable, that the particular confoundings of syntax and semantics in the item forms of Bormuth et al. are only partial. Then, if semantic content is not controlled, spurious measurements of syntax-processing ability may result. If the latter is the case, then a difficulty ordering is premature until a semantic analysis as complete as Bormuth's (1970) syntactic analysis is available. The present results are a demonstration that the difficulty ordering of syntactic structures is not stable, implying that such further analyses are necessary.

There is another problem in measuring syntax difficulty. Sometimes the answer to a question may be betrayed by semantic constraints. The paragraph may contain only one semantically possible answer to the question. For example, a *who* question after a paragraph with only one animate noun could be answered without knowledge of the target's syntactic structure. Such an extreme problem is not likely in a careful study such as that of Bormuth et al. More generally, though, semantic constraints on the answer to a question by the choice of content words for the paragraph are difficult to determine and may not be completely controlled in studies such as that of Bormuth et al. or the present experiment. Again, this may be a moot point if such constraints are perfectly correlated with syntax, but the present results rule out this possibility.

In addition to these two—semantic difficulty and extent to which the paragraph "gives away" the answer—there is a third potential source of variance in the difficulty ordering produced by the Bormuth et al. (1970) method. Two passages that have a target syntactic structure in common may differ in the extent to which they can otherwise be processed to the point at which the critical structure is relevant. One may sensibly hypothesize, for example, that an anaphora cannot be comprehended unless both it and its antecedent are simultaneously in operating (short-term) memory.

Consider the following example sentences:

1. Joe may cry. If *so*, the rest of us will be sad.
2. Joe is coming home. *That* is the best news I've heard all week.
3. Joe may splice the mainbrace. If *so*, the rest of us will be glad.

Bormuth et al. found that passages like Item 1 are comprehended 87% of the time while those like Item 2 are comprehended only 66% of the time, by fourth graders. If he had used a different pro-clause passage like Item 3 instead of Item 1, perhaps only half of the children would have understood pro-clause forms, thus reversing the ranking of clause demonstratives (*that*) and pro-clauses (*so*). Item 3 is longer than Item 1 and may be more difficult to encode due to its idiomatic content. Most important, *so* stands in place of a much more complex construction in Item 3 than in Item 1. Unless a child can record *splice the mainbrace* as a single image, he may lack the ability to process that phrase to the point of being able to resolve the anaphoric reference. No control procedure is available to insure that Bormuth's ranking arises only from syntax differences and not from differences in passage wording or the amount of processing required to get from a syntactic parsing of a sentence to an underlying cognitive representation.

The present results arise from what was expected to be a screening task for an experiment on memory for syntax. Thus, they provide information about only a subset of 9 out of Bormuth et al.'s 14 anaphora forms. The differences between the two studies are as follows: (*a*) the present study used oral, constructed responses while Bormuth et al. used written, multiple-choices responses; (*b*) the present study explicitly controlled the number of semantically plausible potential answers in each passage; (*c*) the location of the target structure in the passage was counterbalanced in the present study; and (*d*) Bormuth et al. used 420 fourth-grade subjects while the present study used 80 subjects from the third and fourth grades.

Method

Subjects

Forty students from a campus laboratory school and 40 students from an urban public school participated as subjects. The campus group ranged in age from 8.0 to 10.0 years with a median age of 8.8 years, while the urban group ranged from 8.0 to 11.5 years with a median age of 9.4 years. All were in third or fourth grade.[4]

Materials

Fourteen anaphora forms were tested in this study. They are listed, with examples, in Table 1. Three items were written for each form. An item consisted of a paragraph plus a question. The paragraph, in turn, consisted of two filler sentences plus a two-sentence critical-structure sequence. The first sentence of the critical sequence contained an antecedent which was referenced by an anaphora in the second sentence. Each of the three items for a form had a different location for the critical sentences in the paragraph: before, between, or after the filler sentences. Each item was constructed so that there were two semantically sensible answers to the question. The correct choice was determined by the anaphora syntax. The question for each item was written by substituting the appropriate *wh-* word for the anaphora and then applying the shortest sequence of transformations that would turn the anaphora sentence into a question.

Each question was typed on a 8½ inch × 11 inch sheet in .42-centimeter gothic type and covered with a clear plastic sleeve. The 42 pages were presented to subjects in a loose-leaf binder. Order of occurrence of the items was approximately counterbalanced.

Procedure

The procedure was first explained to the subject. Each subject worked individually and at his own pace reading each paragraph and then orally answering the question. The subjects from each school were split into two groups of 20 each. One group read the passages (generally silently), while the other group followed along reading (silently) as they listened to a tape recording of the passage. After completing 30 of the 42 passages, each subject stopped for a brief cookie break. Occasionally, the subject would respond to a question by stating the anaphora word (e.g., *he*). The subject was then asked to be more specific (e.g., *Whom do you mean by* he*?*).

Scoring

Each response to a passage by a given subject was punched onto a separate computer card. A list was then produced, sorted by item. This made it possible to examine all answers to an item at once. Each answer was scored as correct only if the referent of the anaphora in the given passage was stated or implied by a stated synonym or superordinate category term. Answers referring to more entities than the correct referent were counted wrong. Scoring was verified by a second observer.

[4] The campus laboratory school actually had two-year groupings, one of which was equivalent to third and fourth grades.

Results

Two reliability measures were computed. The cell means for the 14 anaphora types were compared for the two schools ($r = .94$, $df = 12$, $p < .001$) and for the two presentation conditions ($r = .88$, $df = 12$, $p < .001$); both correlations are quite high, suggesting that these results are relatively stable for the particular items used.

To determine whether the anaphora forms differed in their difficulty, a 20 × 14 × 2 × 2 (Subjects × Anaphora × Schools × Presentation Mode) partially nested analysis of variance was performed on the scores (0–3 correct) of the subjects on each of the anaphora forms. There were significant differences among the anaphora forms ($F = 19.6$, $df = 13/988$, $p < .0001$), but there were no interactions of anaphora with school or presentation mode (Fs ≤ 1.28, $df = 13/988$). The percentage correct for each of the anaphora types is shown in Table 1, and Tukey's (a) post hoc comparisons are shown in Table 2.

Schools was a significant variable ($F = 27.5$, $df = 1/76$, $p < .0001$) with the campus lab school ($M = 2.42$) showing higher performance than the urban public school ($M = 2.04$). There is also some reason to suspect that while the urban school subjects performed equally under the two presentation conditions (Ms = 2.03, 2.04), the campus lab school subjects benefited from the read-and-hear condition (Ms = 2.54, 2.29; $F = 2.89$, $df = 2/76$, $p < .10$).

The important result is the comparison of the rank orders of difficulty for those nine anaphora conditions common to both the present study and to that of Bormuth et al. (1970). The mean percentages correct can be seen in Table 1. There is a significant negative correlation between the two sets of means (Spearman $\rho = .66$, $n = 9$, $p < .05$).

TABLE 1
RELATIVE DIFFICULTY OF ANAPHORIC STRUCTURES

Anaphora type	Example	% correct		
		Bormuth et al.[a]	Present study	Difference
Personal pronoun	Joe left the room. *He* had	64.5	91.7	+27.2
Relative pronoun	The man *who* lives next door	82.8	78.3	−4.5
Noun phrase, demonstrative	The old dog belongs to Joe. *That* is his	81.5	75.0	−6.5
Negated pronoun	The gang went shopping. *No one* bought	67.4	67.5	+0.1
Pro-verb, *so–do*	John likes Mary. *So does* Bill.	82.8	60.8	−22.0
Pro-verb, *so–be/have*	Joe is sick. *So is* Bill.	76.1	54.6	−21.5
Pro-adverb, locative	I am upstairs. It is cold *there.*	—	79.2	—
Pro-clause, *so*	Joe may go. If *so*, we will	86.8	64.2	−22.6
Clause, demonstrative	Joe is dead. *That* leaves two of us.	66.3	90.0	+23.7
Pro-adjective	John is careful. Bill is *that way*, too.	—	66.2	—
Semantic substitute—noun	Those dishes are expensive, but this *china*	65.5	82.9	+17.4
Semantic substitute—verb	Jim shot first. John *fired*, too.	—	82.5	—
Semantic substitute—clause	Bill went to the bank. That *trip* made us short one outfielder.	—	75.0	—
Semantic substitute—adverb	John works carefully. Bill also works *precisely.*	—	70.8	—

[a] Bormuth, Manning, Carr, and Pearson (1970).

DISCUSSION

Since Bormuth et al. (1970) suggest that their difficulty ordering may be the basis for a learning hierarchy for anaphora syntax, it is important to be certain that their data and the present data cannot possibly be a basis for the same hierarchy. The negative rank correlation gives some evidence in this regard. However, it would be possible to find such a correlation if one subject sample possessed the skills needed to comprehend Structure A while a second subject sample had not yet acquired those skills. If both samples performed at a middle level on Structure B, then the first would display a rank ordering of A > B while the second had ordering B > A.

Many structures of language can only be understood in all contexts after a set of several interpretation rules have been learned.

Consider the following example sentences:

4. He had won the race.
5. John knew he had won the race.

For example, a six-year-old knows enough of the rules for understanding personal pronouns to know that in Item 4, *he* refers to the person most recently talked about. However, the child would assume that Item 5 referred to two different people (Chomsky, 1969; Palermo & Molfese, 1972). When two syntactic forms are tested without knowledge of the rules or strategies that different-age children use to comprehend them, it is quite conceivable that their relative difficulty will oscillate as new rules are learned that enable one structure or the other to be understood in more contexts.

This is not the case for the present data since some forms, such as the two pro-verb forms and the pro-clause (*so*) form, are better comprehended by the Bormuth et al. (1970) subjects, while others, such as the personal pronoun, are better comprehended by the present subjects. Since these order differences are absolute rather than relative, we conclude that the Bormuth et al. procedure suffers from a problem more basic than developmental variability.

The present study was intended as a screening task for an experiment derived from Bormuth's work. Thus, every effort was made to reproduce any potentially critical procedures of Bormuth's. Results obtained are statistically reliable. We conclude that syntax is not the basis of a hierarchy of comprehension skills that have not been completely acquired by the time children are in the fourth grade.

TABLE 2

POST HOC COMPARISONS BETWEEN ANAPHORIC STRUCTURES (TUKEY'S *a*)

Anaphora type	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Pro-verb, *so–be/have*				*	*	**	**	**	**	**	**	**	**	**
2. Pro-verb, *so–do*							**	**	**	**	**	**	**	**
3. Pro-clause, *so*									**	**	**	**	**	**
4. Pro-adjective									*	*	**	**	**	**
5. Negated pronoun										*	**	**	**	**
6. Semantic substitute—adverb											*	*	**	**
7. Semantic substitute—clause													**	**
8. Noun phrase, demonstrative													**	**
9. Relative pronoun													*	**
10. Pro-adverb, locative														*
11. Semantic substitute—verb														
12. Semantic substitute—noun														
13. Clause, demonstrative														
14. Personal pronoun														

* $p < .05$.
** $p < .01$.

Bormuth et al. (1970) claimed to have less confidence in their anaphora results because of the multiple-choice procedure used exclusively in that part of their study. Thus, one could argue that what the present paper has really done is to correct a problem in Bormuth's procedure (Bormuth, personal communication, October 1973), thus giving a more correct syntax-learning hierarchy. However, there are very few syntax forms which children in the fourth grade cannot understand in at least some contexts. Several of the "hardest" forms in the Bormuth et al. study are "easy" in the present data.

Why, then, do children perform poorly on a particular structure on even one of the two tests. There are two potential reasons: (*a*) the child may not know the interpretation rules required to understand the structure in a particular semantic context; or (*b*) he may lack the real-time capacity for applying those rules. There is good reason to believe that both of these factors play a role but that the second is of predominant importance for children as old as the present subjects. The available evidence is primarily restricted to studies of pronominalization.

Fredrick, Golub, and Johnson (1970) have found that multiple-choice pronoun items similar to those of Bormuth et al. (1970) are correctly answered only 28% to 50% of the time. The items are grammatically similar to the personal and relative pronoun items of the present study, which are correctly answered 92% and 78% of the time, respectively. The apparent difference is in the level of semantic ambiguity in the two cases. Consider the following examples, from Fredrick et al. and the present study, respectively:

6. The notebook on her desk covered up my drawing *which* was very messy.
   [Identify *which*].
   (a) notebook
   (b) desk
   (c) covered up
   (d) drawing
   (e) messy

7. Two men were walking down the street. One man had on a hat. The big man who was standing on the corner is my father. Who was standing on the corner?

The first example (Item 6) is semantically more complicated and requires that the child realize that the "closest-semantically-acceptable-antecedent" rule applies, even though there are three semantically acceptable potential antecedents, while there are only two potential answers to Item 7. Even children who get Items 6 and 7 correct are quite likely to waver in handling Item 8.

8. I put the package on the table. Because it was tilted, *it* fell off. [Identify *it*.]
   (a) package
   (b) table
   (c) tilted
   (d) fell
   (e) I

The point is that grammatical rules for nine-year-old children are not abstract structures that apply mechanically. They are inextricably bound up with semantics (cf. Palermo & Molfese, 1972). The presence of some pronoun interpretation rules in even five-year-olds (Chomsky, 1969) does not mean that these rules operate abstractly and free of semantic influence. The potential reliability and validity of syntax tests with uncontrolled semantics is low.

The second potential reason for variability in tests of syntactic competence is differences in the extent to which a given item exceeds or stays within the child's channel capacity as an information processor. Recent findings by the present author (Lesgold, 1972) suggest that adult-like comprehension of personal-pronoun sentences is more likely when a child can replace cumbersome surface-structure segments with imaginal codes. Thus, imagery factors apparently play a role in comprehension performance.

These arguments suggest that the design of syntax comprehension curriculum cannot readily be based upon tests of the sort described here and by Bormuth et al. (1970). Such tests have uncontrolled variance due to imagery and semantic factors. The structures tested are often (but not always, cf. Palermo & Molfese, 1972) understood by children when presented in circumstances that have relatively simple semantics. On the other hand, sufficiently complex semantics will probably override knowledge of any structure in young children. Perhaps, semantic analysis at the level of care that Bormuth (1970) has brought to syntactic analysis will be useful in the design of instruction. However, at the present, the use of difficulty orderings for syntax without regard to semantics is not likely to lead to improved instruction in comprehension.

It is certainly possible that abstract syntax rules are being acquired during the ages from which Bormuth's and the present samples were drawn. Further, it is conceivable that one might want to test for the presence of such rules. This could be done by using nonsense words instead of lexical words in the various item forms used in this study. This would yield measures of "pure" syntax ability. However, many syntactic rules interact with semantics, so this pure ability might not be very relevant.

REFERENCES

Bormuth, J. R. *On the theory of achievement test items.* Chicago: University of Chicago Press, 1970.

Bormuth, J. R., Manning, J., Carr, J., & Pearson, D. Children's comprehension of between- and within-sentence syntactic structures. *Journal of Educational Psychology*, 1970, **61**, 349–357.

Carroll, J. B. Defining language comprehension: Some speculations. In J. B. Carroll and R. O. Freedle (Eds.), *Language comprehension and the acquisition of knowledge.* Washington, D.C.: V. H. Winston, 1972.

Chomsky, C. *The acquisition of syntax in children from 5 to 10.* Cambridge: MIT Press, 1969.

Frase, L. T. Maintenance and control in the acquisition of knowledge from written materials. In J. B. Carroll and R. O. Freedle (Eds.), *Language comprehension and the acquisition of knowledge.* Washington, D.C.: V. H. Winston, 1972.

Fredrick, W. C., Golub, L. S., & Johnson, S. L. *Analysis of the Linguistic Ability Test, Grades 4 & 6.* (Tech. Rep. 121) Madison: Wisconsin Research and Development Center for Cognitive Learning, 1970.

Lesgold, A. M. *Effects of pronouns on children's memory or sentences.* (Tech. Rep. 1972/17) Pittsburgh, Penn.: University of Pittsburgh, Learning Research and Development Center, 1972.

Palermo, D. S., & Molfese, D. L. Language acquisition from age five onward. *Psychological Bulletin*, 1972, **78**, 409–428.

(Received May 16, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 339-347

TASK, LEARNER, AND TREATMENT VARIABLES IN INSTRUCTIONAL DESIGN

JOHN E. RHETTS[1]

University of Minnesota

This research studied the effects of three variables (learner, task, and treatment characteristics) on subject performance in two basic learning paradigms (visual discrimination or "matching" and paired-associate tasks). Significant interactions were found for number of errors on both tasks: Learner Characteristic × Task Characteristic on the matching task and Learner Characteristic × Treatment Mode on the paired-associate task. Response latencies and postresponse feedback examination times were analyzed, and generally, both were inversely proportional to the number of errors. Results support the need for instructional research to be based on the Attribute × Treatment interaction model and to employ a "task-first" approach.

The long-standing interest among psychologists in "individual differences" has more recently led to a great deal of speculation and research on reliable patterns of interaction among learner, task, and treatment variables in order to be able to individualize instruction. Cronbach (1957) first gave the name "aptitude-treatment interaction" to such patterns. A basic assumption underlying the aptitude-treatment-interaction concept is that individuals are indeed meaningfully different and that with sufficient conceptual and methodological tools these differences can be significantly utilized in the design of instruction. The aptitude-treatment-interaction believer rejects the notion that there is one best instructional procedure for teaching all individuals.

There are strong arguments in support of significant individual differences from many avenues of research, ranging from behavioral genetics to psychology (McClearn & Meredith, 1966; Newell & Simon, 1972). This position does not deny that there are demonstrable similarities or seemingly universal characteristics among individuals but rather that, *in addition*, individuals vary in many potentially significant ways. The pressing question is to identify and verify patterns of characteristics which are both (*a*) consistently ascribable to or manifested by an individual and (*b*) related to performance differences on some task(s). In such an endeavor, we must attend to the fact that the existence of such patterns is the product of an interaction: one of the organism with the environment. Psychologists have long differed over which side of the interaction merits most attention, debating whether or not the "equipment" the individual carries with him exerts primacy over the structure and content of the task environment. However, the more potentially informative pursuit would be rigorous work on the following questions:

1. For a given task, what identifiable and replicable patterns of characteristics among individuals are associated with significantly different levels of performance on the task?
2. Are such patterns of individual-difference characteristics involved in aptitude-treatment interactions across a variety of tasks or are they task specific?
3. Are performance differences eradicable (partially or completely) by changing or modifying the task environment?

This set of interrelated questions implies an interaction model. Ample support for

[1] Requests for reprints should be sent to John E. Rhetts, Educational Psychology, 330 Burton Hall, University of Minnesota, Minneapolis, Minnesota 55455.

this general approach can be found in research from all the behavioral and social sciences. Vale and Vale (1969) succinctly state three compelling reasons why psychologists must seriously study human performance from an interaction approach:

1. Decades of debate over the so-called nature–nurture controversy have clearly revealed that neither variable can be ignored; interaction studies are more likely to reveal important findings than attempts at hair-splitting over whether organism or environment alone accounts for more of the variance.

2. Psychologists are becoming more interested in the study of processes within individuals and the interface of these information-processing capacities with the task at hand (e.g., Cronbach, 1967; Jenkins, 1967; Newell & Simon, 1972).

3. Only by incorporating individual-difference characteristics as positive elements (as opposed to "error variance") can psychology build adequate general laws about human behavior. Psychologists cannot afford to ignore the demonstrable genetic and behavioral differences among individuals and cannot, therefore, search only for general, invariant functional relations among variables.

Cronbach and Snow (1969) exhaustively reviewed hundreds of published aptitude-treatment-interaction (ATI) studies and concluded that

> Progress toward the goal of identifying and understanding ATI has been slight.... There are no solidly established ATI relations even on a laboratory scale, and no real sign of any hypothesis ready for application and development [p. 193].

Bracht (1970) appears basically to concur with these findings:

> Experimenters usually identified alternative treatments and then through trial-and-error tried to find personological variables to interact with treatments. The analysis of an interaction effect was often an afterthought rather than a carefully planned part of the experiment, i.e., the alternative treatments were not developed with the ATI concept in mind. This approach has not been successful for finding meaningful ... interactions [p. 639].

Rhetts (1972) has proposed a research strategy designed to improve the information yield from aptitude-treatment-interaction studies. First, experimenters must distinguish among three aspects of the aptitude-treatment-interaction design: task characteristics, learner characteristics, and mode of presentation or instruction (treatment) variables. Analysis of published studies (e.g., Bracht, 1970) reveals that aptitude-treatment-interaction researchers have generally conceptualized the aptitude-treatment-interaction problem as two dimensional: learner variables and a diffuse combination of task and treatment variables. However, *both* treatment and task characteristics can be expected potentially to enter (singly and jointly) into interaction with learner characteristics, thus strongly suggesting that a three-dimensional reference scheme should be used.

A second recommendation (Rhetts, 1972) centered upon the sequence of experimental design, the optimum sequence being as follows: first, concentrate on an analysis of the characteristics of the task (and its demands on memory, knowledge, motivation, abilities, etc.); second, identify plausible individual-difference characteristics related to the performance demands of the task; and third, develop different treatments or modes of presentation (different instructions, cues, amount or distribution of practice, available information or materials, etc.) designed to influence performance differences. One reason for adopting a "task-first" sequence is the fact that one of the outstanding features of the human organism is its adaptive capability. This means that the individual can and does modify his behavior contingent upon the situation at hand, by tailoring his own responses to the nature of the task environment confronting him. Thus, while it may be true that an individual manifests a characteristic style or strategy in dealing with some class of tasks or problems, we do not really know enough until we know whether this pattern of similar behavior is (*a*) a response to common stimulus aspects among the tasks or (*b*) an overlearned behavior which is generalized (nonadaptively) to the class of tasks. If *a* is the case, we will also need to know whether the nominal cues or stimuli are effective (salient) for the individual or not and whether the effective stimuli are or

are not relevant to solving the problem or task. By initially attacking the aptitude-treatment-interaction problem at the learner-characteristics dimension, the experimenter cannot know (*a*) what to attribute either the consistency in an individual's behavior or the commonality of behavior among a group of individuals to and (*b*) whether behavioral consistency or commonality is functionally adaptive or nonadaptive (e.g., perfunctorily repetitive). Work by others (e.g., Garner, 1970; Jenkins, 1967; Newell & Simon, 1972) supports a task-first approach.

A second reason for attacking the aptitude-treatment-interaction problem in a task-first sequence is that documented aptitude-treatment interactions could significantly influence and inform the process of instructional design. Such problems begin with the question "what is to be taught" or "what task is the learner to master?" Thus, in this context, it seems inappropriate to begin by asking what learner characteristics or presentational variables are relevant prior to defining the "what" of the instructional problem.

The following research was an effort to investigate interactive effects among learner, task, and treatment variables. The two tasks used were familiar laboratory ones—a visual discrimination or "matching" paradigm and a paired-associate learning paradigm—each with items at two difficulty levels. Two tasks were chosen in order to explore the degree to which performance effects due to impulsivity–reflectivity were either task specific or more generalized. These two particular tasks were chosen both because they appear as ingredients in more complex classroom learning—the ultimate locus of interest—and because an analysis had suggested that their performance demands should differentially affect subjects with different information-processing habits.

The paired-associate paradigm is a two-phase learning task involving both response learning and relationship learning. From an information-processing perspective, this implies that a subject would need to allocate sufficient time to fixate, chunk, and then store the pairs of related information units (first responses, then relations) through short-term memory into long-term memory. Because the items used here were so constructed, the paired-associate task required no extensive search or comparison processing, and given that the response elements (letters) were thoroughly familiar, learner-controlled allocation of time (attention) for fixation and storage was a major task demand. In addition, paired-associate tasks are generally structured so that the subject has access to feedback, and this introduces another control-of-time-allocation variable. As described more fully below, both a self-paced and a fixed-rate, postresponse feedback condition were created. In the self-paced condition, subjects had to make an explicit decision about how long to attend to the feedback information, whereas in the fixed-rate condition, subjects were presented the feedback for an externally controlled interval.

The matching task required subjects to demonstrate their ability to perform perceptual differentiation (Gibson, 1969) among objects (nonsense words) whose features (letters) were themselves already highly learned. This called for scanning and comparing the always-present standard with five alternative responses, identifying the discriminative information locations (distinctive features), and selecting which one of the five response choices was exactly like the standard. This task could be characterized as primarily tapping a performance, in contrast with tapping a learning ability. Given that a subject was thoroughly familiar with the alphabet and had no experimenter-imposed constraints on response or study time, the major demands of this task were on the subject's control of the allocation of attention (time) to processes he/she already was competent in executing: accurate short-term memory fixation and successive scanning and comparing among six items. These operations of scanning for information (discriminating letters) and checking hypotheses about matches, by contrast with the paired-associate task, made no long-term memory demands. Each matching item was a separate event, and the matching items required no response learning and no information from one item

to be remembered or carried forward for use on another item or at a later time.

Undoubtedly, these tasks might be analyzed in other terms and constructs, but the analysis of task demands above led to an interest in the learner characteristic of impulsivity–reflectivity, extensively studied principally by Kagan and his associates (e.g., Kagan, 1965; Kagan et al., 1964). In the present study, impulsivity–reflectivity was conceptualized as a learner characteristic which describes an individual's speed of cognitive information processing. As implied by the task analyses above, on the matching task, speed of processing and responding can be especially crucial at the point of recognizing distinctive or discriminative features and in selecting and evaluating potential correct responses from the array of available responses. On the paired-associate task, this trait should influence the critical function of allocating sufficient time to item fixation and storage into long-term memory of both responses and relations.

In the present study, impulsives were not expected necessarily to perform more poorly (commit more errors) than reflectives. Kagan's classification scheme (impulsive = shorter than median latency *and* greater than median errors) was derived from data on performance tasks of unspecified difficulty. Since number of errors is one index of difficulty, from the data Kagan has published it would appear that his tasks were roughly of moderate difficulty and, therefore, of moderate complexity. Considering the parameters of the human organism as an information-processing system, an individual attempting to process a moderately complex aggregate of information in a short time period could well be expected to make more errors. However, this does not inform us about the relative performance of the two types of individuals at other levels of task complexity. Thus, in this research both easy and difficult items were developed for both tasks.

Hypotheses

This research was designed to test the hypothesis that the learner characteristic of impulsivity versus reflectivity would both generalize to and influence error rates on two basic, learning and performance paradigms (visual differentiation and association learning); that is, this learner characteristic was expected to generalize to more educationally relevant performance situations than the somewhat artificial tasks studied by Kagan and his associates. More specifically, the hypotheses were as follows:

1. Impulsivity–reflectivity would interact with the task characteristic of item complexity such that (*a*) impulsives would perform at least as well as reflectives on low-complexity (easy) items and (*b*) they would perform less well on more complex (difficult) items. Less complex items involve or contain fewer "chunks" of information to be processed, and therefore a rapid responding strategy should not degrade performance on easy items as it would be expected to on more complex ones. The reflectives' longer deliberation should only improve his performance (relative to an impulsive subject) on more complex or difficult items. On a less complex item, added deliberation should have *no* functional value, even possibly resulting in debilitation through confusion or compulsive inability to decide on a response.

2. Impulsivity–reflectivity would interact with the treatment variable of control over potential feedback such that (*a*) impulsives would not self-manage an opportunity to gain corrective feedback so as to optimize their performance as compared with other impulsives in a fixed-pace feedback condition and (*b*) impulsives would make significantly more errors than reflectives in a self-paced feedback condition. This effect was expected to be heightened on complex information items as compared with easy items. In general, impulsives were expected to "rush on" and fail to take advantage of available feedback when they had self-control over the feedback interval. Therefore, they were expected to perform more poorly than reflectives who were expected to take greater advantage of such feedback when they were self-paced. Also, self-paced impulsives were expected to perform more

poorly than impulsives exposed to fixed-pace feedback.[2]

Method

Subjects

A group of 64 second-grade children (24 boys and 40 girls) were administered the matching and paired-associate measures described below. These subjects were identified as the 32 more impulsive and 32 more reflective children from among the entire second-grade population ($n = 90$) of a public elementary school. (The complete testing of all 90 was not possible due to teacher reactions to the multiple absences from class required by the procedure.) The selection of the 64 was made on the basis of the individual's mean response time across the 12 items of the Matching Familiar Figures Test, developed by Kagan et al. (1964).

Stimulus Materials

Two types of performance tasks were used: a matching-with-standard task and a paired-associate learning task. Individual items for the matching task were nonsense words, that is, three- or five-letter groups which did not have any interior vowels. The easier trigrams ($n = 14$) were constructed from low-pronounceability trigrams. The more difficult five-letter ($n = 14$) words were composed by putting together two of the low-pronounceability trigrams and dropping one letter. The difficulty level of the items was systematically varied according to four criteria: first, the number of letters in the item (three or five letters); second, the number of distinctive letters (nonidentical) in the standard as compared with the competing (incorrect) alternative responses (ranging from none to the maximum number of letters in the item); third, the shape of letters (whether ascending, descending, or line letters) as similar or contrasting in the discriminative positions of incorrect alternatives compared with the standard, correct response; fourth, the location in the word of the discriminative letter(s), that is, the first position, last position, middle position, etc. [the first position is the first attended to by beginning readers, the last is next in order, and the middle (or interior) positions are the last to be attended to as discrimination or information locations; e.g., Williams, Blumberg, & Williams, 1970].

The paired-associate items were composed of pairs of alphabetic letters and line drawings representing fragments of the respective letters. The partially isomorphic fragments were the stimuli, and the letters were the responses. Five difficult and five easy items comprised a list. Item difficulty was based on the number of letters which made up the response to a given item (one versus two) and on whether the orientation of the stimulus figure was constant from trial to trial. The more difficult items required a two-letter response, and the orientation (though not the shape) of the stimulus varied from one trial to another. [Rotations of 90, 180, and 270 degrees from the upright (normal) position were used.] Thus, as part of the task for the difficult items, a subject saw four orientations of each stimulus and had to equate them effectively into a single stimulus. On the easy items, a given stimulus always had the same orientation in space and required a one-letter response. Within each difficulty level, four random orders of the five pairs were constructed; these were used to control for serial-learning effects.

Procedure

All material in the matching and paired-associate tasks was presented via slide projector onto a rear projection screen. On the matching task, the subject responded by pressing the one of five buttons which was aligned with the alternative on the screen he selected as being correct. Each subject was instructed to make his choice as quickly as he could and yet get it right and to continue to make a choice on each item until he selected the correct alternative. If he made an incorrect choice, the same slide stayed in view. If he selected correctly, the next slide was presented on the screen.

On the paired-associate task, after one initial exposure to the entire set of stimulus–response pairs on the screen, the subject was instructed to say aloud the name of the letter which "went with" the stimulus figure (when it appeared in view for five seconds) as soon as he could remember or thought that he could. Guessing was stated to be acceptable. The subject was immediately informed about the correctness of his answer both verbally ("right," "wrong") and visually (green light, red light). After each response, the subject was shown the stimulus–response pair. Half the subjects (fixed feedback condition) saw this slide for a fixed, five-second interval, while the other subjects (self-paced feedback condition) could themselves determine how long they looked at the postresponse pair slide.

It was reasoned that perhaps testing all subjects on easy items prior to the difficult ones (or vice versa) might generate a general warm-up effect on

[2] The effect of this treatment variable was studied only on the paired-associate task. The principal reason was that treatment variables (in the three-dimensional aptitude-treatment interaction (ATI) scheme proposed by Rhetts, 1972) are of interest chiefly as instructional manipulations or conditions and in respect to their impact on learning. As described above, the matching task was construed as primarily a performance task. Experimentally manipulating access to feedback in the matching task would have changed its nature from a performance task to one involving learning; it would have introduced a memory demand vis-à-vis response learning before subjects could have begun information search and comparison processing. However, altering this visual differentiation task so that it involved a learning demand could well be a useful experimental manipulation to study in future research.

the items presented second. Also, having the subjects uniformly perform one task prior to the other might introduce some systematic though unpredictable effect due to task order. Hence, for control but not analysis purposes, subjects were randomly assigned to balanced groups with respect both to order of task presentation and order of items worked on first (easy or difficult), thereby tending to spread any such sequence (administrative) effects over all the treatment (experimental) effects (Winer, 1962).

Results

Analysis

Main and interaction effects were examined for the following independent variables:

1. Learner characteristic—impulsivity versus reflectivity;
2. Task characteristic—item difficulty (easy items versus difficult items; within-subjects, repeated-measure variable);
3. Treatment variable—fixed versus self-paced postresponse feedback (for paired-associate task only); and
4. Sex—boys versus girls (analyzed as a control variable).

Separate analyses were done for each of three dependent variables:

1. Number of errors,
2. $\log_e$ (response latency), and
3. $\log_e$ (postresponse, pair-slide examination time) for self-paced subjects on the paired-associate task only.

The number of errors was the dependent variable of primary interest; response and examination times were included in order to assist in interpreting and/or explaining interaction and main effects on the error scores. This yielded a total of five separate analyses (two for the matching task and three for the paired-associate task). In each analysis, the within-subjects error term was used to test all main and interaction effects involving the repeated measures variable of item difficulty; the between-subjects error term was used for all other main and interaction effects.

Findings

Matching task. The number of errors made by the impulsive and reflective subjects on the easy and difficult items of the matching task bears directly on one of the study's hypotheses; namely, rapid processing and responding should not handicap impulsives on the easy items, but the reflectives should make fewer errors on the difficult items. The information in Figure 1 shows that the Learner Characteristic × Task Characteristic interaction conformed to the predicted pattern ($F = 9.77$, $df = 1/60$, $p < .003$). It can be seen that, while all subjects made fewer errors on the easy items, the impulsives made fewer errors than the reflectives. And, although all subjects made substantially and significantly more errors on the difficult items, this increase was greater for the impulsives than the reflectives. Significant main effects were also found for item difficulty ($F = 90.44$, $df = 1/60$, $p < .001$, confirming the a priori classification of items as easy and difficult), for impulsivity–reflectivity ($F = 5.82$, $df = 1/60$, $p < .02$, impulsives more errors overall), and sex ($F = 4.64$, $df = 1/60$, $p < .04$, girls fewer errors overall).

The data for $\log_e$ (response time) on the matching task paralleled those for number of errors. Both impulsives and reflectives took significantly more time to respond to the difficult items than to the easy ones ($F = 294.40$, $df = 1/60$, $p < .001$), and reflectives took significantly more response time overall than did the impulsives ($F = 24.18$, $df = 1/60$, $p < .001$). There was a trend ($F = 3.19$, $df = 1/60$, $p < .08$) for impulsives, in contrast to reflectives, to increase their allocation of time to difficult items compared to easy ones. There was also a trend ($F = 2.84$, $df = 1/60$, $p < .10$) for girls to respond more quickly than boys.

Paired-associate task. The largest influence on paired-associate-task error rates was the difficulty level of the items; overall, the subjects made more than four times as many errors on the difficult items as they did on the easy ones ($F = 200.74$, $df = 1/56$, $p < .001$, as with the matching task, clearly confirming the a priori categorization of items as easy and difficult). The treatment variable of fixed versus self-paced feedback was involved in a three-way interaction ($F = 3.08$, $df = 1/56$, $p < .08$) with both sex and impulsivity–reflec-

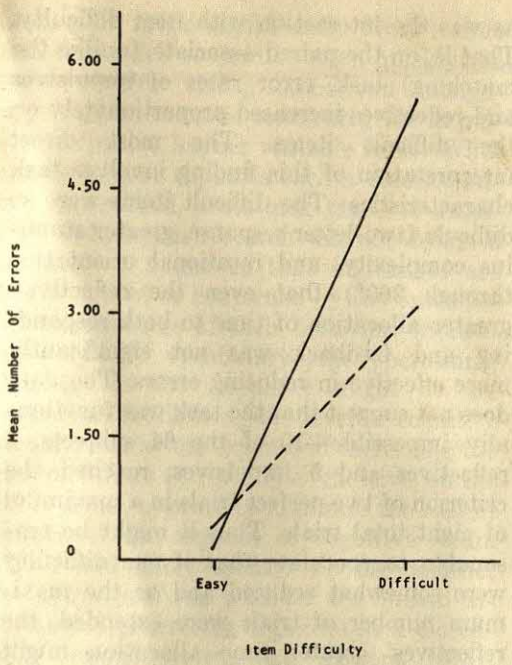


FIGURE 1. Matching task mean number of errors for the Impulsivity–Reflectivity × Item Difficulty interaction.

tivity (see Figure 2). The markedly different performance patterns seen in Figure 2 for impulsives compared with reflectives led to a nested analysis of sex by feedback within the learner characteristic factor (for impulsives, $F = 5.37$, $df = 1/56$, $p < .02$; for reflectives, $F = .03$, $df = 1/56$, $p > .87$). Although the means suggested that (as hypothesized) both impulsives and reflectives did more poorly under the self-paced feedback condition and that the detrimental effect was greater on impulsives than reflectives, the F ratios for feedback ($F = .44$) and learner characteristic by feedback ($F = .08$) were not significant. The hypothesized Learner Characteristic × Item Difficulty interaction did not materialize here as it did on the matching task ($F = .08$).

An analysis of time allocated by "self-paced" subjects to getting feedback revealed the following: reflectives made significantly greater use ($F = 9.47$, $df = 1/28$, $p < .005$) of the feedback opportunity on both easy and difficult items; all subjects paused significantly longer ($F = 29.48$, $df = 1/28$, $p < .001$) on difficult (versus easy) items; and impulsives and reflectives equally increased their use of such an opportunity on difficult items compared with easy items ($F = .22$, $df = 1/28$, $p > .65$). Parenthetically, over all paired-associate items, the self-paced subjects allocated an average of 4.79 seconds to the postresponse pair slide as compared with the experimenter-established interval of 5.0 seconds for subjects in the fixed-pace condition.

The response time data for the paired-associate task generally paralleled those for the matching task: reflectives took significantly longer ($F = 4.96$, $df = 1/56$, $p < .03$) to respond overall; all subjects significantly increased ($F = 148.02$, $df = 1/56$, $p < .001$) their response latency on difficult items; and reflectives did this somewhat more so ($F = 3.23$, $df = 1/56$, $p < .08$) than impulsives on the difficult items.

DISCUSSION

Confirmation was found for each of the study's hypotheses. The characteristic of speed of responding manifested itself on

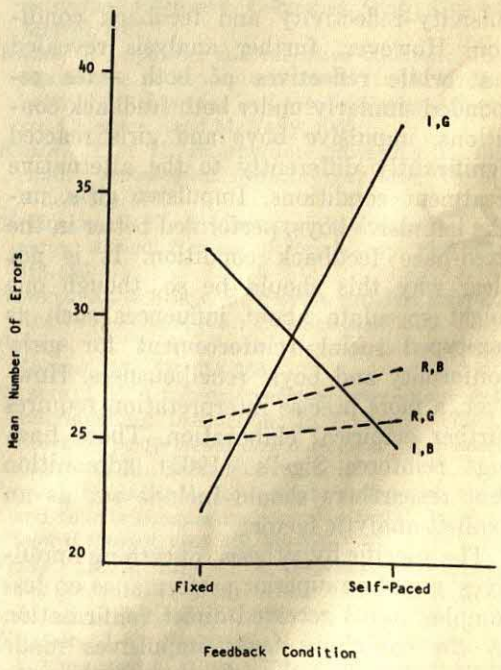


FIGURE 2. Paired-associate task mean number of errors for the Impulsivity–Reflectivity × Sex × Feedback Condition interaction. (Abbreviations: I = impulsives, R = reflectives, B = boys, and G = girls.)

both tasks, with impulsives consistently and significantly showing shorter latencies on each task, both for input (feedback examination) as well as output (response) activities. Thus, this learner-characteristic variable seems to be one which is not task specific. It was also associated with error rates on both tasks: (*a*) It interacted significantly (on the visual discrimination task) with the intratask characteristic of item difficulty. (*b*) It was involved (on the paired-associate task) with both sex and the treatment variable of control over access to feedback in a three-way interaction which did not reach statistical significance.

Overall, the point-biserial correlation between subjects' sex and the impulsivity–reflectivity measure (the Matching Familiar Figures Test) was $-.03$ ($n = 90$). Stratifying the analyses by sex revealed that girls made fewer errors than boys on the matching task while at the same time responding more rapidly. Sex manifested no significant main effect on paired-associate errors, though it did enter into a nearly significant ($p < .08$) three-way interaction with impulsivity–reflectivity and feedback condition. However, further analysis revealed that while reflectives of both sexes responded similarly under both feedback conditions, impulsive boys and girls reacted significantly differently to the alternative treatment conditions. Impulsive girls, unlike impulsive boys, performed better in the fixed-pace feedback condition. It is not clear why this should be so, though one might speculate about influences such as sex-typed social reinforcement for girls' conformity and boys' rebelliousness. However, a more precise interpretation requires further empirical elaboration. These findings reinforce Sigel's (1965) admonition that researchers should include sex as an explicit analytic factor.

The specific hypothesis concerning impulsives' equal or superior performance on less complex items received direct confirmation on the matching task: impulsives made fewer errors than the reflectives on the easy items but made more errors than the reflectives on the difficult items. On the paired-associate task, the main effect for impulsivity–reflectivity was clearly nonsignificant, as was the interaction with item difficulty. That is, on the paired-associate (unlike the matching) task, error rates of impulsives and reflectives increased proportionately on the difficult items. The most direct interpretation of this finding involves task characteristics: The difficult items were so difficult (two-letter response, greater stimulus complexity, and rotational orientation through 360°) that even the reflectives' greater allocation of time to both responding and feedback was not significantly more effective in reducing errors. The data does not suggest that the task was functionally impossible—10 of the 64 subjects, 5 reflectives and 5 impulsives, reached the criterion of two perfect trials in a maximum of eight total trials. Thus it might be reasonable to speculate that if the difficulty were somewhat reduced and/or the maximum number of trials were extended, the reflectives' greater time allocation might pay off in fewer errors. A more complex-interpretation might be sought in differences among intertask characteristics (demands), but additional data would be required to illuminate this line of interpretation.

Hypothesis 2, concerning impulsives' performance under different treatment conditions, received support. The means for all subjects under the two feedback conditions and for impulsives versus reflectives within each feedback condition were in the "right" direction; but despite the rather large mean differences, they cannot be interpreted as strongly confirming Hypothesis 2 since the small F ratios imply that the within-cell variances were relatively large. However, the triple interaction which also involved sex (Figure 2) clearly showed that, as predicted, the treatment condition had marked impact on impulsives' performance while not markedly influencing reflectives' performance and that impulsives' performance differences were also a function of the subjects' sex.

Generally, there was a strong parallel between number of errors and both response and postresponse feedback times: the longer the response latency or examination time, the fewer the errors. The only notable exception to this was that on the matching

task, girls made significantly fewer errors and responded somewhat more quickly. In light of this strong trend, an intriguing question remains concerning the impulsive–reflective learner variable itself: What causes and/or supports such a characteristic, generalized, cross-task trait? A suspicion (which this research was not designed to resolve) might be that impulsives are less aware of or concerned by the outcomes or consequences of their behavior. In Bruner's (1956) vocabulary, the operational strategy of the impulsive might be described as minimizing the level of cognitive strain, even at the expense of increasing the risk of failure (error). The reflective, on the other hand, appears to minimize the risk of failure at the expense of cognitive strain (time and effort spent). To use another metaphor, one can consider the human information-processing organism as a closed-loop control system (e.g. Powers, 1973). When such a system is confronted with a task (an event in the external environment which is sensed and represented internally), it necessarily confronts both (*a*) the probability (very high) of having to make a response and (*b*) the probability (uncertain) that the feedback received from the environment will trigger an internal error signal and require further responses to reachieve a stabilized state. The impulsive could then be described as functioning to minimize the delay in accomplishing the highly certain need to make a response; the reflective, by contrast, functions so as to minimize the uncertain chance of negative environmental feedback.

In summary, this research demonstrates that learner, task, and treatment characteristics can combine in interaction with one another to produce complex performance differences. These findings, in turn, underline the need for investigators interested in instructional design to utilize an Attribute × Treatment interaction design. However, researchers cannot simply "throw together" combinations of the three variables. A task-first approach (as used here) and a two-stage design for data collection (Rhetts, 1972) should improve significantly the yield from such research. The companion (second) stage for this study would be the exploration of treatment variables which could reduce the performance differences reported above.

REFERENCES

Bracht, G. A. Experimental factors related to aptitude-treatment interactions. *Review of Educational Research*, 1970, **40**, 627–645.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. *A study of thinking*. New York: Wiley, 1956.

Cronbach, L. J. The two disciplines of scientific psychology. *American Psychologist*, 1957, **12**, 671–684.

Cronbach, L. J. How can instruction be adapted to individual differences? In R. M. Gagné (Ed.), *Learning and individual differences*. Columbus, Ohio: Charles Merrill, 1967.

Cronbach, L. J., & Snow, R. E. *Individual differences in learning ability as a function of instructional variables*. (U.S. Office of Education Report 4-6-061269-1217) Stanford, Calif.: Stanford University, 1969.

Garner, W. R. The stimulus in information processing. *American Psychologist*, 1970, **25**, 350–358.

Gibson, E. J. *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts, 1969.

Jenkins, J. J. Individual differences in verbal learning. In R. M. Gagné (Ed.), *Learning and individual differences*. Columbus, Ohio: Charles Merrill, 1967.

Kagan, J. Impulsive and reflective children: The significance of conceptual tempo. In J. Krumboltz (Ed.), *Learning and the Educational Process*. Chicago: Rand McNally, 1965.

Kagan, J., et al. Information processing in the child. *Psychological Monographs*, 1964, **78** (1, Whole No. 578).

McClearn, G. E., & Meredith, W. Behavioral genetics. *Annual Review of Psychology*, 1966, **17**, 515–550.

Newell, A., & Simon, H. *Human problem solving*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Powers, W. T. Feedback: Beyond behaviorism. *Science*, 1973, **79**, 351–356.

Rhetts, J. E. Attribute-treatment interactions and individualized instruction. In L. Sperry (Ed.), *Learning performance and individual differences*. Glenview, Ill.: Scott, Foresman, 1972.

Sigel, I. E. Rationale for separate analyses of male and female samples on cognitive tasks. *Psychological Record*, 1965, **15**, 369–376.

Vale, J. R., & Vale, C. A. Individual differences and general laws in psychology: A reconciliation. *American Psychologist*, 1969, **24**, 1093–1108.

Williams, J. P., Blumberg, E. L., & Williams, D. V. Cues used in visual word recognition. *Journal of Educational Psychology*, 1970, **61**, 310–315.

Winer, B. J. *Statistical principles in experimental design*. McGraw-Hill: New York, 1962.

(Received June 7, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 348-353

TEACHER BEHAVIOR ACROSS ABILITY GROUPS: A CONSIDERATION OF THE MEDIATION OF PYGMALION EFFECTS

JUDITH LANDON ALPERT[1]

New York University

Do teachers use more good behaviors with high- or low-reading-ability groups? *Good* teacher behaviors are defined as those teacher behaviors experts judged likely to increase pupil reading performance. Each of 15 second-grade Catholic school teachers' top- and bottom-reading-group sessions were observed and were tape-recorded 3 times. A total of 90 reading group sessions were analyzed (15 Teachers × 2 Reading Groups × 3 Sessions = 90). The difference between good teacher behaviors with the top and bottom reading group was compared by means of the correlated *t* test. It was found that teachers generally treat the 2 reading groups the same with respect to the good behaviors; that is, teachers treat the two ability groups similarly with respect to amount and quality of reading group time, number of reading group materials, and number of good verbal behaviors. Teachers did show preferential treatment to the bottom reading group by placing fewer pupils in that reading group. The findings indicated that teacher behavior may not be adversely affected by teacher expectation.

Rosenthal and Jacobson (1968) hypothesized that teacher expectancies influence pupil growth in achievement. Serious methodological problems (Barber & Silver, 1968; Jensen, 1969; Snow, 1969; Thorndike, 1968; Thorndike, 1969) made their research highly questionable, and replication attempts (Claiborne, 1969; Kester, 1969; Meichenbaum, Bowers, & Ross, 1969) have yielded mixed results.

Recently, attention has begun to focus on the events mediating teacher expectation and pupil performance. The general conclusion of investigators citing differences in pupil ability as a differentiating factor in teacher–pupil interactions is that low-ability pupils have fewer opportunities to respond in the classroom (Good, 1970; Rist, 1970; Willis, 1970) and receive fewer positive and more negative teacher contacts (Brophy & Good, 1970; Good, 1970; Rist, 1970). Rasmussen's (1961) and Schwarz' (1967) findings, however, do not support the latter conclusion.

One major criticism of most of the investigations concerning teacher expectation and teacher behavior is that observers knew which pupils were of high and low ability. Good's (1970) attempt to control observer bias is an exception. Good did not inform observers about pupil ability. However, it is likely that pupil ability became apparent during classroom observation. The present study investigated the mediation of Pygmalion effects and attempted to control observer bias.

Method

Subjects

The research was carried out in 15 second-grade classes housed in 11 New York City Catholic schools which serve a middle-class population. Schools were selected in which second-grade pupils were grouped into classes randomly and into reading groups by ability. Fourteen of the 15 teachers were lay teachers. Twelve of the 15 teachers held bachelor's degrees.

Instrument

Good teacher behaviors are defined as those teacher behaviors which experts judged likely to

[1] Requests for reprints should be sent to Judith Landon Alpert, Department of Educational Psychology, School of Education, New York University, Washington Square, New York, New York 10003.

increase pupil reading performance. The 39 experts who participated in developing the list included 3 teachers of reading teachers, 15 remedial reading teachers, and 21 first-grade teachers. All experts had served in their position for at least two years.

Since a purpose of the research was to determine if the teacher gives more good behavior to one ability group, an attempt was made to develop a list of good teacher behaviors for the top reading group and for the bottom reading group. However, according to expert opinion, good teacher behaviors do not differ for the two reading ability groups; that is, the experts identified a list of five teacher behaviors which they judged likely to increase both top- and bottom-reading-group pupil performance. The good behaviors are (*a*) more reading group time, (*b*) best reading group time, (*c*) more materials in reading group, (*d*) fewer pupils in reading group, and (*e*) more "good" verbal behaviors. There are 20 good verbal behaviors.

Development

The list was developed as follows.[2] Three reading specialists familiar with the purpose of the study listed four teacher/nonverbal behaviors, each of which they judged to be objectively measurable and likely to increase pupil reading performance. The nonverbal behaviors were listed after discussion, by concensus, and were judged good for both reading groups. Responses to a questionnaire indicated that a minimum of 30 of the 36 reading specialists and teachers agreed that an increase in each of the four nonverbal behaviors was likely to increase the reading performance of pupils in high reading groups and in low reading groups.

The procedure for listing good verbal behaviors was an adaptation of a method used by Ryans (1960) which was based on Flanagan's (1954) critical incidents technique. This procedure involved the following: 36 reading specialists and teachers listed from a tape recording of a teacher who was teaching reading those teacher behaviors likely to increase bottom-reading-group pupil performance and top-reading-group pupil performance. Approximately half of the reading specialists and teachers were asked first to list good teaching for the top reading group. Reading specialists and teachers were also asked to justify their responses. The 20 verbal behaviors were developed from those responses listed by a minimum of 21 of the 36 reading specialists and teachers which could be objectively defined.

Responses indicated that good verbal behaviors were the same for both reading groups. Less than 10 responses were listed as good for one reading group only. An attempt was made to validate the verbal items; that is, three reading specialists were asked to rank order tape recordings of four teachers. The two teachers ranked "less good" by all three specialists used relatively few, good verbal behaviors in comparison to the two teachers ranked "good." Percentage agreement was at least 76% between coders on three consecutive tape recordings of reading group sessions. On one consistency check during the four weeks of coding, the percentage agreement (76%) was maintained.

Operational Definitions

Reading group time was defined as that time period beginning when the teacher called the reading group together and ending when the teacher dismissed the reading group and was no longer working exclusively with any pupils in that reading group. Reading group time did not include time in which the teacher worked with pupils who were not members of the reading group in session. *Best* reading group time was defined as that one third of the morning school session and that one third of the afternoon school session in which the teacher reported she felt most motivated to teach. Reading group materials were defined as those materials which the teacher used for instructional purposes during reading group time. Examples of reading group materials are as follows: blackboard, picture, slide, film, object, text, workbook, teacher-made cards, flashcards, and supplementary reading materials. Reading group pupils were defined as those pupils whose names appeared on the teacher's top- or bottom-reading-group list at mid-year. Good verbal/reading group behaviors were defined as those 20 verbal behaviors listed in Table 2[3] which the teacher verbalized during reading group time.

Procedure

The classes were each visited 4 times over a four-week period. There was a total of 60 visits (15 teachers × 4 visits = 60 visits). The data collectors (15 undergraduate students), teachers, and principals were told that the purpose of the study was to learn more about the learning patterns of pupils of different abilities. The tape recording of reading group sessions enabled the observers to be naive.

The purpose of the first visit was to habituate pupils and teachers to the presence of the observer and the tape recorder. The purpose of the next three visits was to collect data on teacher/nonverbal behavior and to tape-record teacher/verbal behavior.

Data Analysis

Data for teacher/nonverbal behavior were coded from forms completed by data collectors. One exception was the number of pupils, which was determined from the teacher's list at mid-year. Data for teacher/verbal behavior were coded by two coders from tape recordings. Observer bias was minimized by tape-recording sessions, numbering the recordings rather than identifying the reading

[2] A more detailed account can be found in Alpert (1973).

[3] A more detailed account can be found in Alpert (1973).

group on the recording, and organizing them so that the two ability groups taught by one teacher were not coded consecutively. Also, the coders did not tape-record the sessions.

Data for four of the five good behaviors (reading group time, best reading group time, reading group materials, and good verbal behaviors) were analyzed similarly as follows. Sums across the three sessions for each of the good behaviors and for each of the 20 good verbal behaviors were obtained for each teacher's top reading group and bottom reading group. The sums were divided by the number of teachers to obtain means for the two reading groups. The differences between these means were compared by way of the correlated t test. The procedure used to analyze the data for the number of pupils differed in that means indicated the average number of pupils for the two reading groups rather than the average number of pupils across three sessions for the two reading groups; that is, summing across sessions was inappropriate since the number of pupils per reading group was determined from the teacher's list at mid-year rather than from observations or tape recordings.

Results

The correlated t test results in Table 1 indicate that significantly fewer pupils were placed in the bottom than in the top reading group. Only an average of 8.87 pupils were in the bottom reading group as compared with an average of 13.87 pupils in the top reading group. Although the direction of difference between means for the four remaining good behaviors (amount of time, amount of best time, number of materials, number of good verbal behaviors) indicates preferential treatment to the bottom reading group, none of these differences are significant.

TABLE 1

Means, Standard Deviations of Difference, and Correlated t Values for Differences in Good Teacher Behaviors with Top and Bottom Reading Groups

Behavior	M for group		Correlated t	
	Top	Bottom	SD of difference	t
Amount of time	72.27	79.47	15.64	−1.78
Amount of best time	22.60	27.13	34.86	−.50
Number of materials	6.13	6.73	2.13	−1.09
Number of pupils	13.87	8.87	4.83	4.07*
Number of good verbal behaviors	115.87	143.00	66.72	−1.34

* $p < .01$.

TABLE 2

Means, Standard Deviations of Difference, and Correlated t Values for Differences in Good Verbal Behaviors with Top and Bottom Reading Groups

Verbal behavior	M for group		Correlated t	
	Top	Bottom	SD of difference	t
1. Praise reading	9.93	17.13	10.44	−2.67
2. Support	4.93	10.13	8.64	−2.33
3. Reinforce	21.60	30.90	23.50	−1.53
4. Encourage	7.90	12.80	8.17	−2.12
5. Praise behavior	.53	.13	.83	1.87
6. Sequence	1.80	3.07	3.63	−1.35
7. Demonstration	1.33	1.00	2.06	.62
8. Experience	9.80	10.73	5.44	−.66
9. Definition	13.27	11.27	2.78	2.78
10. Comparison	4.47	3.07	3.62	1.50
11. Summarize	2.60	4.07	2.59	−2.20
12. Specific question	18.27	19.70	14.14	−.40
13. Word family	3.47	1.80	5.68	1.14
14. Simplify reading	2.33	3.67	2.89	−1.79
15. Table of contents	.67	.33	.82	1.60
16. Resource	2.60	2.42	3.98	.26
17. Question all	7.87	11.60	8.05	−1.79
18. Concern	.60	.53	1.62	.17
19. Teacher answer	.80	.73	1.16	.23
20. Encourage question	.13	.47	.90	−1.46

Note. The .05 level of significance could not be used for each test since 20 t tests were done. Therefore, a critical region ($CR = 3.65$) was determined from Bonferroni's t statistic with simultaneous .05 significance level under 20 behaviors (Miller, 1967).

Table 2 presents the 20 good verbal behaviors given to the top reading group and to the bottom reading group. Results presented in this table indicate no significant difference in the amount of good behavior given to the two groups. In summary, the data in Tables 1 and 2 indicate that teachers in this study generally treated the two reading groups equally. The small amount of teacher preferential treatment found was directed to the bottom reading group. Since there were fewer pupils in the bottom reading group, each pupil in the bottom group received more good behaviors; that is, each pupil in the bottom group was given more time and more best time, and more good verbal behaviors were directed to each pupil in that group.

Discussion

The results indicate that teachers generally treat the top and bottom reading groups the same with respect to the good behaviors. Several findings in the present study are in accordance with the findings of other investigations. The finding that teachers place significantly more pupils in the top than in the bottom reading group has been reported by Groff (1962). Second, Rasmussen's (1961) and Schwarz' (1967) finding that teachers do not give more supportive comments to high- or low-ability pupils also supports the present finding. Third, Brophy and Good (1970) present data which indicate that teachers spend as much time with pupils in the two reading ability groups.

However, the conclusion of some investigators (Brophy & Good, 1970; Good, 1970; Rist, 1970; Rubovits & Maehr, 1971; Willis, 1970) considering teacher instructional behavior with pupils differing in ability is that high-ability pupils receive preferential treatment. Also, the results from the present study could be interpreted as indicating preferential treatment to the high-ability group; that is, Strong, McCullough, and Traxler (1967) and Spache and Spache (1969) contend that bottom-reading-group pupils need more good teaching than top-reading-group pupils. Since top-reading-group pupils in Grade 2 have basic reading skills, they are more capable of independent work than their less achieving peers. In that light, the general lack of preferential treatment could be interpreted as discriminatory to slow readers.

What is clear, however, is the discrepancy in results between previous investigations and the present one. Moreover, this discrepancy occurred despite a similarity in the process variables investigated; that is, some process variables claimed to be mediators of the self-fulfilling prophecy by previous investigators were also considered in the present study. Specifically, Brophy and Good (1970) found that teachers reinforce and demand quality performance more with the top reading group. Moreover, this occurred when differences in number of correct responses and errors made by pupils were taken into account. Although the teacher behaviors studied by Brophy and Good were defined and scored differently than in the present study, "reinforce quality performance" is similar to a category used in the present study (Verbal Behavior 1, praise). Also, "demanding quality performance" is similar to a combination of two categories used in the present study, (Verbal Behavior 2, support; Verbal Behavior 4, encourage). However, the results in the present study indicate no differential treatment between groups for any of these verbal behaviors. Moreover, when frequencies in Categories 2 and 4 are combined, the difference between groups is significant at the .05 level, indicating more good behaviors directed to the bottom reading group.

Also, Good (1970), Rist (1970), and Rubovits and Maehr (1971) consider praise, and Good (1970), Rist (1970), Willis (1970), and Rubovits and Maehr (1971) consider attention. In the present study, praise was directly measured (Verbal Behavior 1, praise reading; Verbal Behavior 5, praise behavior), and attention may be considered to be indirectly measured (Nonverbal Behavior 1, amount of time). The results in most of the studies considering praise and attention, unlike the present investigation, indicate discriminatory teacher behavior to low-ability pupils.

One explanation for the discrepancy in findings is that Catholic school teachers may be more responsive to the needs of slow learners. However, Catholic school teachers in the present investigation are similar to public school teachers with respect to lay status and education. Difference in control for observer bias is another explanation for the discrepancy in findings. In the present investigation, there was some control for observer bias in coding the verbal behaviors. As indicated, the sessions were tape recorded, the level of the reading group was not labeled on the recording, and the two ability groups taught by one teacher were not coded consecutively. Also, the coders did not tape-record the sessions. Moreover, it is doubtful that observer bias could influence the objective to code nonverbal behaviors.

Observer bias was probably, however, not completely controlled. To consider the control for observer bias in the present study, two first-grade teachers were asked to identify the reading group recorded on randomly selected tapes. Although one teacher identified 6 out of 10 correctly and the other identified 3 out of 6 correctly, it is possible that identification would have improved over trials.

The results of this study indicate a more tentative acceptance of the findings concerning the mediation of Pygmalion effects, a more sophisticated control of observer bias in future studies, and a more rigorous consideration of the needs of high- and low-ability pupils. This conclusion is further substantiated by the paucity of data concerning the relationship between the process variables and pupil performance. That is, the relationship between the process variables and pupil performance has not been demonstrated by investigators considering the mediation of Pygmalion effects (Brophy & Good, 1970; Good, 1970; Rasmussen, 1961; Rubovits & Maehr, 1971; Schwarz, 1967; Willis, 1970) and is not indicated by a review of the literature on teacher behavior and student achievement (Rosenshine, 1971).

Also, the results indicate the need to consider reasons other than teacher discriminatory behavior for the relatively slow growth of low-ability pupils. An alternative explanation is that educators do not know how to instruct pupils of low ability. There is little empirical data indicating appropriate pedagogical strategies for the reading ability groups. Moreover, Chall (1966) indicates that pupils of high and average ability are less affected by teaching methods than low-ability pupils. The development of the list of good teaching behavior afforded a preliminary delineation of appropriate pedagogical strategies for reading ability groups. Nevertheless, the relationship between the good behaviors and pupil performance has not been demonstrated. Whether those teacher behaviors identified by experts or others not yet identified by them affect bottom-reading group pupil performance is an important area for future investigation.

REFERENCES

Alpert, J. L. The effect of a change in teacher behavior on pupil performance. (Doctoral dissertation, Columbia University) Ann Arbor, Mich.: University Microfilms, 1973, No. 73-28, 181.

Barber, T. X., & Silver, M. J. Fact, fiction, and the experimenter bias effect. *Psychological Bulletin*, 1968, **70**(6, Pt. 2).

Brophy, J. E., & Good, T. L. Teachers' communications of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology*, 1970, **61**, 365–374.

Chall, Jeanne. *Learning to read*. New York: McGraw-Hill, 1966.

Claiborne, W. L. Expectancy effects in the classroom: A failure to replicate. *Journal of Educational Psychology*, 1969, **60**, 377–383.

Flanagan, J. C. The critical incidents technique. *Psychological Bulletin*, 1954, **70**, 75–96.

Good, T. L. Which pupils do teachers call on? *Elementary School Journal*, 1970, **70**, 190–198.

Groff, P. J. A survey of basal reading group practices. *Reading Teacher*, 1962, **15**, 232–238.

Jensen, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 1969, **39**, 1–123.

Kester, S. W. The communication of teacher expectations and their effects on the achievement and attitudes of secondary school pupils. Unpublished doctoral dissertation, University of Oklahoma, 1969.

Meichenbaum, D. H., Bowers, K. S., & Ross, R. R. A behavioral analysis of teacher expectancy effect. *Journal of Personality and Social Psychology*, 1969, **13**, 306–316.

Miller, R. *Simultaneous statistical inference*. New York: McGraw-Hill, 1967.

Rasmussen, R. T. A study of differences in the character of the learning environment provided by teachers of high and low ability groups with implications for the development of individual talent. Unpublished doctoral dissertation, Teachers College, Columbia University, 1961.

Rist, R. C. Student social class and teacher expectations. *Harvard Educational Review*, 1970, **40**, 411–451.

Rosenshine, B. *Teaching behaviors and student achievement*. Great Britain: King, Thorne & Stace, 1971.

Rosenthal, R., & Jacobson, L. *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston, 1968.

Rubovits, R., & Maehr, M. Pygmalion analyzed: Toward an explanation of the Rosenthal-Jacobson findings. *Journal of Personality and Social Psychology*, 1971, **19**, 197–203.

Ryans, D. G. *Characteristics of teachers*. Washington, D. C.: American Council on Education, 1960.

Schwarz, M. E. The effect of teacher approval on self-concept and achievement of fourth, fifth and sixth grade children. Unpublished doctoral dis-

sertation, University of Nebraska Teachers College, 1967.

SNOW, R. E. Unfinished Pygmalion. *Contemporary Psychology*, 1969, **14**, 197–199.

SPACHE, G. D., & SPACHE, E. G. *Reading in the elementary schools*. Boston: Allyn & Bacon, 1969.

STRONG, R., MCCULLOUGH, C., & TRAXLER, A. E. *The improvement of reading*. New York: McGraw-Hill, 1967.

THORNDIKE, R. L. Review of R. Rosenthal and L. Jacobson, *Pygmalion in the Classroom. American Educational Research Journal*, 1968, **5**, 708–711.

THORNDIKE, R. L. But you have to know how to tell time. *American Educational Research Journal*, 1969, **6**, 692.

WILLIS, R. J. The influence of teacher expectations on teacher's classroom interaction with selected children. Unpublished doctoral dissertation, George Peabody College for Teachers, 1970.

(Received June 11, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 354-362

TYPE AND FREQUENCY OF QUESTIONS IN PROCESSING TEXTUAL MATERIAL[1]

JOHN P. RICKARDS[2]
Purdue University

FRANCIS J. Di VESTA
Pennsylvania State University

While reading a prose passage, 80 college sophomores answered questions after every two or every four paragraphs of text. Moreover, the subjects responded to meaningful-learning questions requiring subsumption of facts under given ideas, questions involving rote learning of facts, questions demanding rote learning of ideas, or task-irrelevant questions. When questions occurred more frequently, meaningful-learning questions resulted in recall of relevant and incidental information that was equal to or greater than rote-learning-of-ideas or task-irrelevant questions. Only meaningful-learning questions were adversely affected by less frequent pacing. It is argued that the more thorough processing associated with meaningful-learning questions necessitated their relative closeness in text. Furthermore, it appears that different types of questions can influence the direction (forward versus backward) and kind of processing activities associated with them.

Some investigators (Frase, 1968d; McConkie, Rayner, & Wilson, 1973; Rothkopf & Bisbicos, 1967; Watts & Anderson, 1971) of mathemagenic behavior in prose contexts have examined the effects on retention of various kinds of adjunct questions. In all of these experiments, different kinds of questions were found to differentially influence the information readers obtained from the passages they had read.

The main focus of the present investigation is on a comparison between essentially two types of questions. These two types of questions reflect the distinction made by some psychologists (Ausubel, 1968; Di Vesta, 1972) among various levels of learning such as that between rote, verbatim learning and higher order, meaningful learning. In rote learning, material appears to be learned in a relatively random and unordered fashion. More importantly, it is not highly coded, since the subject acquires what is to be learned is exactly the same form as it is presented to him. In the context of the present experiment, rote learning implies that each sentence is processed as a discrete unit with the subject making no attempt to interrelate the various sentences of the passage. It seemed to us that this kind of processing activity might be induced by interspersing questions in a text that required the subject to recall as literally as possible a word or words representing a fact or an idea contained in one sentence of the passage. The verbatim nature of the answer to the question combined with the attentional focus on *one* sentence should effect the desired result, rote learning.

Meaningful learning, on the other hand, involves the organization of facts under given, higher level, related ideas (Ausubel, 1968). Hence, we reasoned that questions requiring subjects to organize or subsume facts in a passage under superordinate ideas provided in the questions would be likely to produce meaningful learning of the questioned material.

Research employing "advance organizers" or themes (Ausubel, 1960; Dooling &

[1] This study is based on a portion of the senior author's dissertation which was submitted to the Department of Educational Psychology, Pennsylvania State University, in parital fulfillment of the requirements for the PhD degree. Francis J. Di Vesta served as chairman of the dissertation committee.

[2] Requests for reprints should be sent to John P. Rickards, Educational Psychology and Research Section, SCC-G #55, Purdue University, West Lafayette, Indiana 47907.

Lachman, 1971) rather than questions as "orienting directions" (Frase, 1970) has demonstrated the superiority of meaningful learning over rote learning of prose material. In light of the above evidence, our primary hypothesis was that meaningful-learning postquestions would facilitate retention more than rote-learning postquestions.

Closely related to the use of questions for controlling mathemagenic behavior is the frequency with which they are employed. It has been theorized (Rothkopf, 1963) that postquestions serve, at least in part, to reinforce appropriate mathemagenic activity. Moreover, it is nearly axiomatic that frequent reinforcement fosters the development of the reinforced behavior. Thus, we predicted that the frequency of the questions would be directly related to performance. This hypothesis has been confirmed in studies by Frase (1968a, 1968c).

Method

Subjects

The subjects were 80 college sophomores enrolled in an introductory educational psychology course. Although participation in the experiment was voluntary, subjects did receive credit toward their course grade for such participation. Each subject was assigned to one of the conditions ($n = 10$) by reference to a table of random numbers.

Materials

The experimental materials were a modification of those used by Bruning (1970) and consisted of an 800-word passage describing the characteristics of a fictitious African nation, "Mala." The various paragraphs of the passage were unrelated to one another to the extent that each paragraph addressed a different aspect of the geography, economy, government, history, or social conditions of the fictitious nation. Each paragraph consisted of a topic sentence (general idea) followed by three related subordinate sentences (specific facts). In accordance with Bruning's (1970) procedure, the allowable relationships between superordinate (topic) and subordinate sentences were as follows:

> (a) the subordinate statement restated the relationship of the response term to the stem portion of the superordinate sentence in more specific and explicit form, (b) the subordinate statement was a specific reason for the relationship between the response term and the stem portion of the superordinate statement, or (c) the subordinate statement revealed a particular characteristic of the response term in the superordinate statement, in some way typifying or describing the nature of the response term itself [p. 188].

The 800-word passage consisted of eight text segments, each of which contained two 50-word paragraphs. For purposes of illustration, one text segment follows:

> The southern area of Mala can best be described as a desert. Rainfall is less than 2 inches per year in southern Mala. The soils in the southern area of Mala are either rocky or sandy. In the summertime temperatures have been recorded as high as 135 degrees in southern Mala.
>
> The history of Mala has been marked by exploitation. The first slaves were forcefully taken from Mala to Europe in 1610. When Europeans came over to Mala to settle there, they never paid the Malans for the land they occupied. Prior to the coming of the Europeans, Arab nomads frequently plundered villages in Mala.

One paragraph from each text segment was always the target for the various types of interspersed questions employed. Such questioned paragraphs were referred to as *related paragraphs;* those not questioned were called *unrelated paragraphs.* Three types of experimental questions were constructed for each of the eight related paragraphs. The first type of question, *rote-learning-of-facts questions,* were directed toward specific facts in the subordinate sentences of the related paragraphs. Literal, verbatim recall of this information was all that was required to answer rote-learning-of-facts questions. Referring to the text segment illustrated above, a rote-learning-of-facts question was "How many inches of rain fall per year in southern Mala?"

A second type of question, *rote-learning-of-ideas questions,* were aimed at ideas represented in the superordinate (topic) sentences of the related paragraphs. Like the rote-learning-of-facts questions, rote-learning-of-ideas questions only required the subjects to literally recall passage material. The rote-learning-of-ideas question appropriate for the related paragraph of the above text segment is "What geographical term best describes southern Mala?"

The third type of experimental question, *meaningful-learning questions,* required the subjects to organize or subsume the facts of the subordinate sentences under the ideas contained in the superordinate topic sentences of the related paragraphs. Using the text segment illustrated above, the meaningful-learning question was "Why can it be said that southern Mala is a desert?" In addition to these three types of experimental questions, control group questions were developed which were totally irrelevant to the passage. Such questions, which were termed *task-irrelevant questions,* were devised to control for whatever effect (e.g., increasing attention) the mere answering of questions might have and were totally unrelated to the passage so as to prevent interference.

A booklet was prepared for each subject containing one text segment per 5 × 8 inch page with separate sheets of paper containing one or two test questions per page.

Design

Each subject responded to one of the four types of questions explicated above, rote-learning-of-facts, rote-learning-of-ideas, meaningful-learning, or task-irrelevant questions. In addition, the frequency of the questions varied such that half of the subjects received *one* question after every two paragraphs of text, while the other half received *two* questions after every four paragraphs of text. All subjects were given the same number of questions; only the frequency of the questions was allowed to vary. Thus, the two between-subjects factors of the design for this study were the level of question variable (meaningful-learning, rote-learning-of-ideas, rote-learning-of-facts and task-irrelevant questions) and the frequency of question variable (after two paragraphs and after four paragraphs).

At the conclusion of the experiment, a free-recall test was administered. Regardless of the particular condition in which a subject participated, each protocol was scored for each of the following dependent variables:

1. *Relevant facts*—facts in the related paragraphs questioned by rote-learning-of-facts questions;
2. *Relevant ideas*—ideas in the topic sentences of the related paragraphs questioned by rote-learning-of-ideas questions;
3. *Relevant subsumed facts*—facts in the related paragraphs other than the relevant facts about which the subject was questioned in the meaningful-learning question condition. In actuality, meaningful-learning questions directed the subject toward *all* the facts (including relevant facts) in the related paragraphs;
4. *Incidental facts*—facts in the unrelated paragraphs, that is, paragraphs about which there were no questions; and,
5. *Incidental ideas*—ideas in the topic sentences of the unrelated paragraphs.

Procedure

Instructions informed the subjects that the purpose of the experiment was to examine the effects on learning of asking questions during the reading of written materials. The instructions also indicated that in answering the questions the subjects would neither be allowed to take any notes nor to turn back to a page once it had been read. The subjects were given 30 seconds to read each page of the passage and 10 seconds to answer each accompanying question. At the conclusion of the experiment, the subjects were given 15 minutes to write down what they could recall about the country of Mala. The entire experiment including the free-recall test required about 28 minutes to administer.

As suggested earlier, each sentence of the passage consisted of either a fact or an idea. For scoring purposes, each sentence was reduced to a phrase of three to four words. This procedure is based in part on a method developed by Cofer (1941) and more recently used by Howe (1970). For example, one of the text sentences reads "In the summertime temperatures have been recorded as high as 135 degrees in southern Mala," and the reduction reads "temperatures 135 degrees in southern Mala." Moreover, alterations in phrasing and synonyms were considered correct if they did not change the meaning of a given sentence. An acceptable semantic equivalent of the text sentence given above might be, "It is very hot in southern Mala."

In order to determine the objectivity of the scoring procedure, two raters judged both the number of ideas and the number of facts generated in each of 45 randomly selected free-recall protocols. The Pearson product-moment correlation for the total number of facts recalled was .96; for the total number of ideas recalled by the subjects, the product-moment correlation was .95. These data provide evidence that the measures were adequately objective for the purposes of the present experiment.

Results

In this experiment, there were three mutually exclusive measures of relevant information: relevant facts, relevant subsumed facts, and relevant ideas. Further, there were two mutually exclusive measures of incidental information—incidental facts and incidental ideas. Accordingly, six separate analyses of variance were done—one for each of the above dependent variables and one for the total amount of recall. Table 1 presents the means for each of the eight conditions of the experiment for each of the dependent variables employed.

Total Recall

The effect of the various types of questions on total recall of information in the passage was significant ($F = 18.12$, $df = 3/72$, $p < .01$) as was the effect due to question frequency ($F = 9.39$, $df = 3/72$, $p < .01$). However, these two variables—question type and question frequency—interacted significantly ($F = 7.81$, $df = 3/72$, $p < .01$).

For more frequent questions (after every two paragraphs), the analyses (Newman-Keuls) demonstrated that meaningful-learning questions produced significantly ($p < .01$) more total recall than any other

TABLE 1

MEANS OF THE EIGHT CONDITIONS OF THE EXPERIMENT FOR EACH DEPENDENT VARIABLE

Dependent measure	Frequency × Level of Question							
	After two paragraphs				After four paragraphs			
	ML	RLI	RLF	I	ML	RLI	RLF	I
Total recall	34.90	27.00	20.90	15.00	20.30	24.40	21.90	15.30
Relevant facts	4.70	2.60	3.20	.90	2.50	3.20	4.00	1.20
Relevant subsumed facts	10.40	8.90	6.00	3.90	6.40	7.20	7.00	4.20
Relevant ideas	5.70	5.10	2.60	2.20	4.10	4.90	2.50	2.20
Incidental facts	10.00	6.60	6.90	4.90	5.30	5.80	5.80	4.90
Incidental ideas	4.10	3.80	2.20	3.10	2.00	3.30	2.60	2.80

Note. Abbreviations: ML = meaningful learning, RLI = rote learning of ideas, RLF = rote learning of facts, and I = task irrelevant.

type question used in this study. Moreover, rote-learning-of-ideas questions yielded greater ($p < .05$) total recall than either rote-learning-of-facts or task-irrelevant questions which, in turn, did not differ from each other. When questions occurred after every four paragraphs, however, only those receiving rote-learning-of-ideas and rote-learning-of-facts questions performed significantly ($p < .05$) better than the control (task-irrelevant) question group.

And finally, question frequency (after two or after four paragraphs) only had a significant ($p < .01$) influence on meaningful-learning questions.

Relevant Facts

Although the effect of the various types of questions on the recall of relevant facts was significant ($F = 15.06$, $df = 3/72$, $p < .01$), the results must be qualified by the significant interaction between the types of questions and their frequency ($F = 5.08$, $df = 3/72$, $p < .01$).

Simple effects of the types of questions for each level of question frequency (Newman-Keuls procedure) indicated that all task-relevant questions (rote-learning-of-facts, rote-learning-of-ideas, and meaningful-learning questions) led to significantly ($p < .05$) better recall than the control (task-irrelevant) questions, regardless of question frequency. Among the contrasts for the three experimental question types, a significant ($p < .01$) difference was found only between meaningful-learning and rote-learning-of-ideas questions, which differed only when questions were presented more frequently (after every two paragraphs). Thus, subjects responding to factual questions did not recall any more of the facts for which they were specifically questioned than did subjects receiving either type of general idea question.

Comparisons (t tests) of the simple effects of question frequency for each type of question revealed that only the difference between meaningful-learning questions presented after two paragraphs and after four paragraphs of text was significant ($t = 2.99$, $p < .01$). Thus, only subjects responding to meaningful-learning questions performed better when the questions occurred more frequently.

Relevant Ideas

In the analysis of relevant ideas, only the effect due to the types of questions was significant ($F = 23.14$, $df = 3/72$, $p < .01$). Using the Newman-Keuls procedure, both the effects of meaningful-learning questions ($\bar{X} = 4.90$) and rote-learning-of-ideas questions ($\bar{X} = 5.00$) were found to be significantly ($p < .01$) greater than rote-learning-of-facts questions ($\bar{X} = 2.55$) or task-irrelevant questions ($\bar{X} = 2.20$). No other mean differences were significant ($p > .05$) in this analysis. These results imply that questions requiring either meaningful learning or rote learning of ideas lead to greater

recall of the questioned ideas than questions requiring common knowledge or rote learning of facts. Further, with these question types, frequency does not appear to be an influential factor in the recall of the questioned ideas.

Relevant Subsumed Facts

The effect of the various types of questions on the recall of relevant subsumed facts was significant ($F = 11.44$, $df = 3/72$, $p < .01$). However, this variable interacted with question frequency ($F = 3.68$, $df = 3/72$, $p < .01$).

For questions occurring more frequently (after every two parapraphs), analysis using the Newman-Keuls procedure revealed that idea-oriented questions (meaningful learning and rote learning of ideas) were significantly ($p < .05$) superior to questions demanding rote learning of facts or questions testing common knowledge (task-irrelevant questions). In this case, superiority was measured in terms of the number of facts recalled that were tapped *only* by the questions (meaningful learning) requiring subsumption of those facts. Thus, focusing attention on a superordinate idea in a paragraph rather than on one of the subordinate facts yields greater recall of the remaining facts in a paragraph.

The only significant simple effect due to question frequency was that for meaningful-learning questions ($t = 3.42$, $p < .01$). Comparable to the results presented for relevant facts, meaningful-learning questions were more effective when questions appeared relatively close together; rote-learning questions, on the other hand, were unaffected by question frequency.

Incidental Facts

Analysis of the data for recall of incidental facts revealed effects for question type ($F = 3.58$, $df = 3/72$, $p < .05$), question frequency ($F = 7.71$, $df = 1/72$, $p < .01$), and the Type of Question × Frequency interaction ($F = 3.08$, $df = 3/72$, $p < .05$).

For questions following every two paragraphs of the passage, Newman-Keuls comparisons indicated that meaningful-learning questions were significantly ($p < .05$) greater in the recall of incidental facts than rote-learning-of-ideas, rote-learning-of-facts, or task-irrelevant questions, which did not differ significantly from each other. No differences ($p > .05$) among contrasts of treatment means were found when questions were used after every four paragraphs of text. The analyses of the two levels of question frequency for each of the four types of questions yielded only a significant simple effect due to frequency for meaningful-learning questions ($t = 3.96$, $p < .001$). Again, these data indicate that the effect of questions requiring meaningful learning of material is apparent only when such questions occur relatively frequently.

Incidental Ideas

The analysis of the scores based on the recall of incidental ideas demonstrated an effect due to the frequency of the questions ($F = 4.29$, $df = 1/72$, $p < .05$) and a Type of Question × Frequency interaction ($F = 3.06$, $df = 3/72$, $p < .05$).

The analyses of simple effects via the Newman-Keuls procedure demonstrated that when questions were used more frequently, both meaningful-learning and rote-learning-of-ideas questions led to significantly ($p < .05$) better recall than that of the rote-learning-of-facts questions but not of the task-irrelevant questions. No other mean differences were significant for either level of question frequency. And again, the only significant simple effect due to question frequency was that for meaningful-learning questions ($t = 3.48$, $p < .001$).

Correlational Analyses

Correlations among the five dependent variables were computed for each of the conditions employing different types of questions presented after every two paragraphs of text. These four conditions (Meaningful Learning 2, Rote Learning of Ideas 2, Rote Learning of Facts 2, and Task Irrelevant 2) consistently yielded different amounts of recall as demonstrated in all of the above analyses. We reasoned that correlational analyses of the sort mentioned above might aid us in our interpretation of these effects.

The results revealed two significant correlations for Rote-Learning-of-Ideas 2 questions and four significant correlations for Meaningful-Learning 2 questions. No significant correlations among the dependent variables were found for Rote-Learning-of-Facts 2 or Task-Irrelevant 2 questions.

Regarding Rote-Learning-of-Ideas 2 questions, a significant correlation ($r = .67$, $p < .05$) was observed between relevant ideas and incidental ideas, thereby suggesting that Rote-Learning-of-Ideas 2 questions may have engendered a set to learn ideas. Among subjects given Rote-Learning-of-Ideas 2 questions, there was also a tendency to recall together facts from the related paragraphs ($r = .70$, $p < .05$, between relevant facts and relevant subsumed facts). Taken together, these correlations suggest that Rote-Learning-of-Ideas 2 questions produced a tendency for facts to become associated with other facts (at least in related paragraphs) and ideas to become associated with other ideas, but *not* a tendency for facts to become associated with ideas.

While this latter-mentioned tendency did not occur with Rote-Learning-of-Ideas 2 questions or with any other type of question, it was manifest to a significant degree with Meaningful-Learning 2 questions ($r = .74$, $p < .01$, between relevant ideas and relevant facts; $r = .62$, $p < .05$, for relevant ideas and relevant subsumed facts; $r = .74$, $p < .01$, between incidental ideas and incidental facts). These data strongly suggest that Meaningful-Learning 2 questions engendered subsumption, the process whereby facts are related to ideas.

The final correlation ($r = .64$, $p < .05$, between relevant facts and incidental facts) implies that if subjects given Meaningful-Learning 2 questions recalled relevant facts, they also tended to recall incidental facts.

Recall in Related and Unrelated Paragraphs

A separate analysis of variance was performed for Meaningful-Learning 2, Rote-Learning-of-Ideas 2, and Task-Irrelevant 2 questions using one within-subjects factor consisting of the following four levels: (*a*) relevant ideas, (*b*) total related facts (relevant facts and relevant subsumed facts combined), (*c*) incidental ideas, and (*d*) incidental facts. Since the total number of facts from the related and the unrelated paragraphs was slightly different (29 for related, 26 for unrelated), the data were converted into percentages for these analyses.

All three questions types (Meaningful Learning 2, Rote Learning of Ideas 2, and Task Irrelevant 2) yielded a significant effect due to the repeated measures ($F = 13.66$, $p < .01$; $F = 15.69$, $p < .01$; $F = 5.86$, $p < .01$, respectively).

Further analyses (Newman-Keuls procedure) demonstrated that while the Task-Irrelevant 2 questions did not differ significantly in the recall of information (facts or ideas) from the related or the unrelated paragraphs, both the Meaningful-Learning 2 and the Rote-Learning-of-Ideas 2 questions did produce significantly ($p < .05$) more information (either of facts or ideas) from the related paragraphs than from the unrelated paragraphs. These data suggest that subjects given either Meaningful-Learning 2 or Rote-Learning-of-Ideas 2 questions reviewed the material for answers to the questions posed (see Discussion section).

Discussion

Forward- and Backward-Processing Effects

In concert with previous investigators (e.g., Rothkopf & Bisbicos, 1967), the present results clearly indicate that text-relevant postquestions can produce a substantial amount of recall of information that is both questioned and not questioned. In the present study, such effects were found with both meaningful-learning and rote-learning-of-ideas questions when placed after every two paragraphs of text (Meaningful Learning 2 and Rote Learning of Ideas 2, respectively). What processing activities produced these effects?

Frase (1967) originally proposed that the learner activities associated with adjunct postquestions involved either a forward process (shaping or eliciting appropriate

reading skills) or a backward process (reviewing of previously read material for answers to questions posed). More recently, Frase (1968b) has indicated that both forward and backward processing probably occur with adjunct postquestions. There is some evidence (McGaw & Grotelueschen, 1972) to support this contention at least with respect to verbatim postquestions (of the fill-in type, demanding literal recall of facts). Boyd's (1973) data, however, suggest that verbatim postquestions similarly affect retention and rehearsal (backward process) but not attention (forward process).

The separation in this experiment of text segments into related (questioned) and unrelated (nonquestioned) paragraphs allowed a distinction to be made between forward and backward effects.

A forward effect for the experimental questions would result in more recall than control (task-irrelevant) questions from both the related and the unrelated paragraphs. By operating in a forward manner, the experimental postquestions would be irrelevant to both paragraphs in text segments following them, and so any treatment effects due to these questions would be equally present in both kinds of paragraphs. On the other hand, because task-irrelevant questions are substantively irrelevant to all paragraphs in the passage, they should result in equally low performance on the recall of material from either type of paragraph. A backward effect for the experimental postquestions would result in more recall than the task-irrelevant questions in the related paragraphs but not necessarily in the unrelated paragraphs. In this case, the experimental postquestions would engender review of previously read material including the material actually questioned (specific backward effect) and perhaps information adjacent to and/or thematically related to the material questioned (general backward effect).

Since Rote-Learning-of-Facts 2 questions only produced significantly more questioned facts (relevant facts) than the control questions, we can conclude that only a specific backward effect was produced by Rote-Learning-of-Facts 2 questions.

Meaningful-Learning 2 questions, however, produced more recall than task-irrelevant questions in both the related and the unrelated paragraphs. Evidently, Meaningful-Learning 2 questions influenced processing behaviors related to attention (forward effect). The high correlation for Meaningful-Learning 2 questions between recall of facts and ideas from both types of paragraphs ($r = .66$, for related paragraphs; $r = .74$, for unrelated paragraphs) strongly suggests that the type of attentional process present was a set to subsume facts under given ideas.

Moreover, there is evidence in this study which seems to indicate that Meaningful-Learning 2 questions also produced a backward effect. Substantially more facts and ideas were recalled in the related than in the unrelated paragraphs with Meaningful-Learning 2 questions but not with task-irrelevant questions. It appears, therefore, that Meaningful-Learning 2 questions engendered both a forward effect (set to subsume) and a backward effect (review of questioned material).

Relative to control questions, Rote-Learning-of-Ideas 2 questions yielded more recall of facts and ideas in the related paragraphs but not in the unrelated paragraphs. Also, significantly more information (facts or ideas) was recalled in the related than in the unrelated paragraphs. Thus, we can assume that a backward effect was produced by Rote-Learning-of-Ideas 2 questions. In this regard, it is interesting to note that while facts in the related paragraphs tended to be recalled together ($r = .70$), ideas and facts in these paragraphs were not so recalled ($r = .30$). In view of the above correlational data, it is quite unlikely that the backward process involved relating facts to ideas, as in subsumption. Perhaps, then, in the course of answering questions about the ideas in the related paragraphs, subjects given Rote-Learning-of-Ideas 2 questions reviewed or rehearsed facts in the paragraphs containing the questioned ideas. By dint of this review, these rehearsed facts

tended to become associated together at recall.

There is also some evidence of a forward effect operating with Rote-Learning-of-Ideas 2 questions. Significantly more ideas in the related paragraphs were recalled in the Rote-Learning-of-Ideas 2 than in the Task-Irrelevant 2 question condition. While not significant ($p > .05$), Rote-Learning-of-Ideas 2 questions tended to produce more recall of ideas in the unrelated paragraphs as well (Rote Learning of Ideas 2 = 3.8; Task Irrelevant 2 = 3.1). Furthermore, with Rote-Learning-of-Ideas 2 questions, there was a significant relationship between recall of ideas in the related and the unrelated paragraphs ($r = .70$ for Rote Learning of Ideas 2; $r = .03$ for Task Irrelevant 2). It seems that we might cautiously interpret these findings to mean that Rote-Learning-of-Ideas 2 questions produced a set to learn ideas (a forward effect).

In comparing Meaningful-Learning 2 and Rote-Learning-of-Ideas 2 questions, it is interesting to note that while both types of questions yielded the same number of *ideas* from either the related or the unrelated paragraphs, Meaningful-Learning 2 questions were significantly superior to Rote-Learning-of-Ideas 2 questions in terms of the recall of facts from either type of paragraph. As mentioned previously, with Meaningful-Learning 2 questions, ideas and facts in either the related or the unrelated paragraphs tended significantly to be recalled together ($r = .66$ and $r = .74$, respectively). This tendency was not observed with Rote-Learning-of-Ideas 2 questions ($r = .30$ and $r = -.03$, respectively). The above evidence seems to suggest that Meaningful-Learning 2 questions engendered a subsumptive process, whereas no such processing occurred in the case of Rote-Learning-of-Ideas 2 questions. Thus, even though subjects provided with Meaningful-Learning 2 and Rote-Learning-of-Ideas 2 questions had the same number of ideas available at recall, the prior subsumption of the material associated with Meaningful-Learning 2 questions greatly facilitated the recall of the more detailed factual material. Using lists of words, Tulving and Pearlstone (1966) found results analogous to those above in which a prose passage was employed. More specifically, they found that superordinate category names (potential subsumers) were more effective in producing the subordinate words when present *both* during learning (presumably engendering a subsumptive process) and recall than at recall alone.

In summary, it appears that different types of adjunct postquestions produced processing behaviors that varied in both direction and kind. Rote-Learning-of-Facts 2 questions only produced a specific review (a specific backward effect) of the material questioned. Rote-Learning-of-Ideas 2 questions, however, were evidently responsible for a general review (general backward effect) of material thematically related and adjacent to the ideas questioned. There is also some evidence that Rote-Learning-of-Ideas 2 questions produced a set to learn ideas (a forward effect). Finally, Meaningful-Learning 2 questions apparently engendered both a set to subsume (forward effect) and a review (backward effect) process.

Frequency of Different Type Questions

Unlike rote-learning questions (rote learning of facts and rote learning of ideas), meaningful-learning questions were influenced by question frequency. Also, only Meaningful-Learning 2 questions produced a significant amount of recall of material from both the related and the unrelated paragraphs. Apparently, these questions induced more thorough processing of the passage than did either of the rote-learning questions. Thus, it seems that placing meaningful-learning questions relatively far apart may have overtaxed the subjects' processing capacity, that is, produced excessive "cognitive strain" (Bruner, Goodnow, & Austin, 1956), thereby eliminating their advantage over the other types of questions. On the other hand, placing verbatim-learning questions close together or for apart resulted in very little difference in cognitive strain, since the processing be-

haviors associated with these question types were less extensive.

Educational Implications

Questions were effectively used in this experiment to induce meaningful learning, thereby supporting and extending the work of Ausubel and his colleagues (see Ausubel, 1968) who have employed another form of orienting direction, advance organizers, for this purpose. Furthermore, the present study supports and extends the findings of research conducted by Rothkopf and Bisbicos (1967) and Frase (1968c). More specifically, adjunct postquestions were found to influence incidental learning. Furthermore, the results imply that meaningful learning questions are superior to factual (rote-learning-of-facts) questions, even when the amount of time spent reading the passage is the same for subjects responding to either type of question. In order to be effective, however, meaningful-learning questions must be spaced relatively close together, so as to minimize cognitive strain. In certain situations, such as the one found in the present study as well as in Boyd's (1973) experiment, factual postquestions have only direct instructive effects, that is, they produce learning of question-relevant facts, but they do not produce incidental learning of facts or general ideas. Under the same conditions, however, the present experiment suggests that meaningful-learning questions induce processing behaviors that favorably influence the recall of both relevant and incidental material. Since meaningful-learning questions contribute to the acquisition of ideas as well as facts, it appears entirely reasonable to conclude that the material is learned in an organized rather than a discrete fashion—a most significant "fringe benefit" of meaningful learning.

REFERENCES

AUSUBEL, D. P. The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of Educational Psychology*, 1960, **51**, 267–272.

AUSUBEL, D. P. *Educational psychology: A cognitive view*. New York: Holt, Rinehart & Winston, 1968.

BOYD, W. M. Repeating questions in prose learning. *Journal of Educational Psychology*, 1973, **64**, 31–38.

BRUNER, J. L., GOODNOW, J. J., & AUSTIN, G. A. *A study of thinking*. New York: Wiley, 1956.

BRUNING, R. Short-term retention of specific factual information in prose contexts of varying organization and length. *Journal of Educational Psychology*, 1970, **61**, 186–192.

COFER, C. N. A comparison of logical and verbatim learning of prose passages of different lengths. *American Journal of Psychology*, 1941, **54**, 1–21.

DI VESTA, F. J. An evolving theory of instruction. *Educational Technology*, 1972, **12**, 34–39.

DOOLING, D. J., & LACHMAN, R. Effects of comprehension on retention of prose. *Journal of Experimental Psychology*, 1971, **88**, 216–222.

FRASE, L. T. Learning from prose material: Length of passage, knowledge of results and position of questions. *Journal of Educational Psychology*, 1967, **58**, 266–272.

FRASE, L. T. Effect of question location, pacing and mode upon retention of prose material. *Journal of Educational Psychology*, 1968, **59**, 244–249. (a)

FRASE, L. T. Questions as aids to reading: Some research and theory. *American Educational Research Journal*, 1968, **5**, 319–332. (b)

FRASE, L. T. Some data concerning the mathemagenic hypothesis. *American Educational Research Journal*, 1968, **5**, 181–189. (c)

FRASE, L. T. Some unpredicted effects of different questions upon learning from connected discourse. *Journal of Educational Psychology*, 1968, **59**, 197–202. (d)

FRASE, L. T. Boundary conditions for mathemagenic behaviors. *Review of Educational Research*, 1970, **40**, 337–347.

HOWE, M. J. A. Repeated presentation and recall of meaningful prose. *Journal of Educational Psychology*, 1970, **61**, 214–219.

MCCONKIE, G. W., RAYNER, K., & WILSON, S. Experimental manipulation of reading strategies. *Journal of Educational Psychology*, 1973, **65**, 1–8.

MCGAW, B., & GROTELUESCHEN, A. Direction of the effect of questions in prose material. *Journal of Educational Psychology*, 1972, **63**, 580–588.

ROTHKOPF, E. Z. Some conjectives about inspection behavior in learning from written discourse and the response mode problem in programmed self-instruction. *Journal of Programmed Instruction*, 1963, **2**, 31–46.

ROTHKOPF, E. Z., & BISBICOS, E. Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology*, 1967, **58**, 56–61.

TULVING, E., & PEARLSTONE, Z. Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, 1966, **5**, 381–391.

WATTS, G., & ANDERSON, R. C. Effects of three type of inserted questions on learning from prose. *Journal of Educational Psychology*, 1971, **62**, 387–394.

(Received June 13, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 363-366

SUBJECT AREAS AND COGNITIVE PRESS

JOE M. STEELE
University of Colorado

HERBERT J. WALBERG[1]
University of Illinois at Chicago Circle

ERNEST R. HOUSE
University of Illinois at Urbana

From 121 Illinois classes in four basic subject areas, Grades 6 through 12, student perceptions were obtained on the emphasis given to cognitive objectives derived from the six levels of the Bloom taxonomy. Using covariance control for class size, grade level, and class type, discriminant analysis revealed three significant (probability less than .001) functions: the first, here termed "convergence–divergence" distinguishes mathematics from language arts classes; the second, "syntax–substance" separates language arts and mathematics from science and social studies; and the third, "objectivity–subjectivity" distinguishes science from social studies. The findings together with those from a prior study, if further replicated, may suggest objective empirical ways of identifying the structure of curriculum and instruction.

The curriculum has been logically divided into subject matter areas on a priori grounds since ancient times. The continuing influence of Aristotle, Comte, and other early deductive theorists is reflected in most contemporary work in curriculum development (Ford & Pugno, 1964). In the last decade, psychologists have carried out several empirical studies to explore the structure of the curriculum. The present paper extends this work by determining the emphasis given to various cognitive objectives (Bloom, 1956) as perceived by classes in several subject areas.

In the earliest attempt to identify variables that distinquish the press on students in various major fields of study in higher education, Thistlethwaite (1962) analyzed 10 measures of student perception on the Inventory of College Characteristics. Physical sciences and mathematics students perceived strong press for Scientism, Compliance, and Vocationalism but weak press for Humanism and Independence. Humanities and social studies students perceived strong press for Humanism, Independence, and Enthusiasm but weak press for Scientism, Compliance, and Vocationalism.

Astin (1965) suggested that the student perceptions of classroom environment are a useful basis for classifying different subject fields empirically. A factor analysis of college-student questionnaire responses produced three factors. Factor I was Foreign Language versus Social Science, with Foreign Language characterized by enthusiastic instructors who knew their students by name while Social Science was characterized by little classroom discussion, little homework, and arguing with the instructor. Factor II was Natural Science versus English and Fine Arts, with the former high on students not speaking in class and the latter high on class discussion, humor, and diverse opinions. Factor III was Business versus History, with more testing emphasis, less research emphasis, and duller instructors in Business.

In a semantic-differential, factor-analytic study of sixth- through ninth-grade classes, Yamamoto, Thomas, and Karns (1969) found that mathematics and science classes rated high on a factor labelled "Vigor" (alive–large–strong–fast). Social sciences and language arts rated high on a "Certainty" factor (safe–easy–usual–familiar).

[1] Requests for reprints should be sent to Herbert J. Walberg, College of Education, University of Illinois at Chicago Circle, Box 4348, Chicago, Illinois 60680.

Anderson (1971) compared student perceptions of the social environment of 54 science, mathematics, French, and humanities classes in eight Montreal high schools. A multivariate analysis of covariance tested the relation of the 15 Learning Environment Inventory scales to the subject matter classifications with class size, sex of students, and IQ as covariates; discriminant analysis was also employed. The first discriminant function (see Figure 1) separated mathematics from science, French, and humanities (English plus history) classes; mathematics classes were characterized as being high in friction and in favoritism but low on formality. The second function separated science classes seen as disorganized and satisfying from humanities classes seen as apathetic. The third function separated French from the other subjects with French seen as goal directed and dissatisfying.

Differences in criteria characterizing subject areas, as well as different ways of grouping areas for analysis, made summary of the foregoing studies difficult. Clearly, however, students perceive subject matter areas differently. The purpose of the present study was to examine the differences among four subject areas on student perception of the emphasis given to various cognitive processes. In an exploratory study such as this, it seemed reasonable to concentrate on subject areas generally considered "core" or basic subjects. The value of using subject areas that allow at least a partial replication of previous work (Anderson, 1971) also supported the choice of subject areas.

Method

Data were collected from 121 independently constituted classes in 69 Illinois schools as part of a statewide evaluation of gifted programs (Steele, House, & Kerins, 1971). Over 20 classes representing each of the four basic subject areas of science, mathematics, social studies, and language arts, and ranging from Grades 6 to 12 composed the sample. About one third of the classes in each subject area were composed of students identified as gifted.

Twenty-three items from the Class Activities Questionnaire were used to obtain the students' perception of prevailing patterns of instructional emphasis. Students were asked to agree or disagree on a 4-point scale to items describing the general kinds of activities which characterized their class. These items were derived from six levels of cognition on the Bloom (1956) *Taxonomy* and several affective conditions that might be stressed in class (see Table 1, which contains the items with significant loadings on discriminant functions). A description of the development of this instrument and evidence for reliability are presented in Steele, House, and Kerins (1971).

Analysis

A multivariate analysis of covariance with discriminant functions was carried out on the class means of the items. To introduce as much statistical control as possible, several possibly contaminating effects were removed from the criteria by covariance adjustments before testing the differences among the four subject matter areas. The covariates included the following variables: linear—grade level, class size, type (coded 1 if designated gifted, 0 otherwise); quadratic—grade level squared and class size squared; and interactions—Grade Level × Class Size, Grade Level × Type, and Class × Type.

Results and Discussion

Table 1 shows the magnitude and significance of the discriminant variates: the first discriminant function accounts for 69% of the variance; the second function, 17%; and the third, 13%. The chi-square approximations to Wilk's lambda (Bock, 1963) for all three functions are highly significant (probability less than .001). Figure 1 shows the positions of subjects on the discriminant functions and Anderson's (1971) Montreal functions with the zero points set at science in both cases for comparability.

"Loadings," that is, the correlations of the criterion variables with the discriminant functions, were calculated for each Illinois function. Fifteen of the 23 Class Activities Questionnaire items had significant loadings (probability less than .01) on the first function. Six items had significant loadings on the second function and 8 items had significant loadings on the third function.[2] The items with significant loadings on the first function, which is termed "convergence–divergence" and which contrasts language arts and mathematics (see Figure 1) sug-

[2] Interested readers can write the authors for a copy of the complete questionnaire and loadings.

TABLE 1

VARIANCE ACCOUNTED FOR AND SIGNIFICANCE OF DISCRIMINANT FUNCTIONS

Discriminant function	Discriminant variance (%)	Chi-square	*df*
1. Convergence–divergence	69.12	262.78*	69
2. Syntax–substance	17.50	114.07*	44
3. Objectivity–subjectivity	13.38	51.24*	21

* Significant beyond the .001 level.

gest the two well-established psychometric factors—verbal and number. The variables associated with the language arts side of the dimension have to do with the cognitive operations: interpretation (Items 16 and 6), evaluation (Item 20), synthesis (Item 23), and translation (Item 9). Affective processes include student independence (Item 14) and participation in discussion (Item 5). The items associated with this pole of the discriminant function all suggest an openness to or consideration of many alternative "answers" or the creation of new ones. Associated with mathematics are variables dealing with analysis (Items 12 and 7), memory (Items 1 and 10), the affective conditions of test and grade stress (Items 8 and 22), the absence of humor (Item 25), and little discussion (Item 15). A reading of these items reveals the strong flavor of converging on a single "right" answer in a methodical, no-nonsense way. While the naming of functions is in the last analysis an intuitive act, the tenor of the items themselves supports the labels selected. Figure 1 shows that in the first and largest discriminant functions of the Illinois and Montreal studies, mathematics is sharply distinguished from other subjects and appears to be seen as unpleasant. The similarity of these findings, despite the use of different instruments, is striking.

Loadings on the second function, termed "syntax–substance," represent a contrast of language arts and mathematics with science and social studies. The function appears to represent Schwab's distinction of tool and application subjects in the curriculum. The variables associated with mathematics and language arts include the cogni-

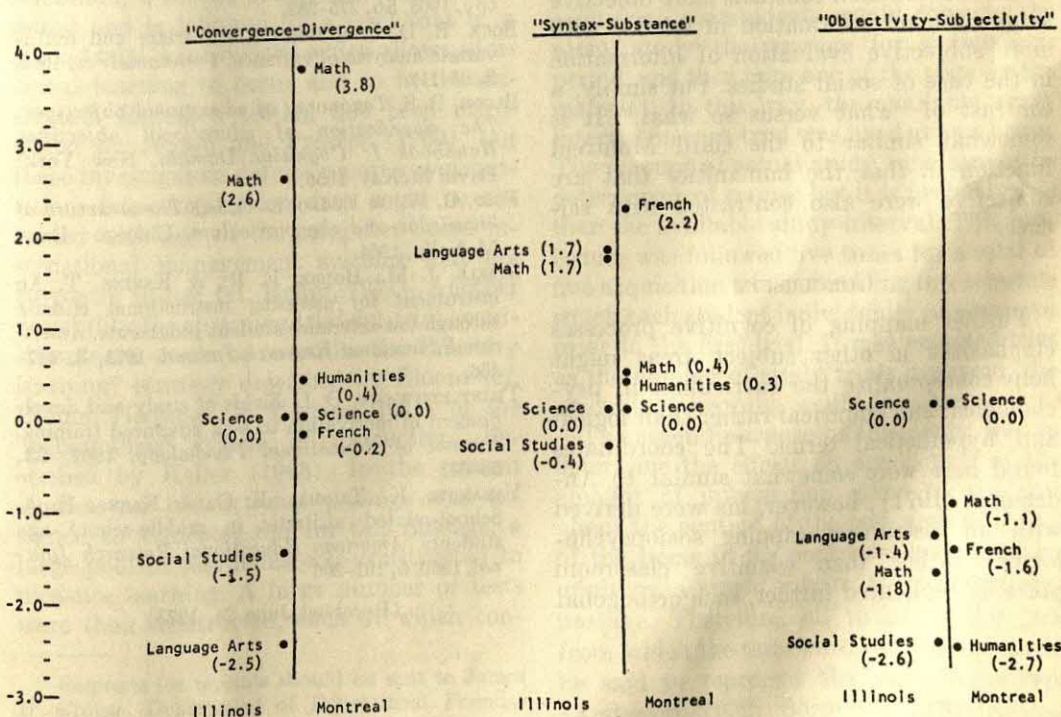


FIGURE 1. Subjects in discriminant space.

tive operations: synthesis (Items 23 and 11), translation (Item 9), and application (Item 3). Associated with science and social studies are the cognitive processes of summarizing (Item 21), memorizing (Items 1 and 10), and evaluating (Item 20). The second function in the Montreal study is similar to this function in that the study of French is more syntactical than substantive.

Loadings shown on the third discriminant function termed "objectivity–subjectivity" contrast science with social studies. The variables associated with science include independent exploration (Item 14), learning and memorizing (Items 1, 8, and 10), interpreting (Item 6), and synthesizing (Item 23). Associated with social studies are variables dealing with evaluation (Items 2 and 20) and the absence of humor (Item 25). This function might well have been labeled "doing versus judging." The items, however, suggest an orientation to the seeking and processing of external data as opposed to reviewing information and its implications in the light of a personal set of values. Thus, this function contrasts more objective acquisition of information in science with more subjective evaluation of information in the case of social studies. Put simply, a contrast of "what versus so what." It is somewhat similar to the third Montreal function in that the humanities that are subjective were also contrasted with science.

Conclusion

Further mapping of cognitive processes emphasized in other subject areas might help conceptualize the curriculum in psychological and empirical rather than logical and hypothetical terms. The coordinates found here were somewhat similar to Anderson's (1971); however, his were derived with an instrument tapping sociopsychological rather than cognitive classroom press. If replicated further, such orthogonal coordinates may organize the relations of the subject areas in objective, interpretable, parsimonious ways. Since cognitive processes and aspects of the social environment transcend the curriculum, designing instruments to measure them which focus on the press or context of instruction may enable the further relating of psychological processes to subject areas in ways not ordinarily thought of by curriculum makers, teachers, and students. For example, the present research raises such questions as, Is mathematics inherently convergent? If not, why is it apparently conveyed convergently to the student? Should the syntactical aspects of science and social studies be given more emphasis? Can some subjective aspects of science and some objective qualities of social studies and humanities be imparted?

References

Anderson, G. J. Effects of course content and teacher sex on the social climate of learning. *American Educational Research Journal*, 1971, **7**, 649–664.

Astin, A. W. Classroom environment in different fields of study. *Journal of Educational Psychology*, 1965, **56**, 275–282.

Bock, R. D. Programming univariate and multivariate analysis of variance. *Technometrics*, 1963, **5**, 95–117.

Bloom, B. S. *Taxonomy of educational objectives: The classification of educational objectives. Handbook I: Cognitive Domain.* New York: David McKay, 1956.

Ford, G. W., & Pugno, L. (Eds.) *The structure of knowledge and the curriculum.* Chicago: Rand McNally, 1964.

Steele, J. M., House, E. R., & Kerins, T. An instrument for assessing instructional climate through low-inference student judgments. *American Educational Research Journal*, 1971, **8**, 447–466.

Thistlethwaite, D. L. Fields of study and development of motivation to seek advanced training. *Journal of Educational Psychology*, 1962, **53**, 53–64.

Yamamoto, K., Thomas, E. C., & Karns, E. A. School-related attitudes in middle-school age students. *American Educational Research Journal*, 1969, **6**, 191–206.

(Received June 24, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 367-372

ACQUISITION OF COLLEGE COURSE MATERIAL UNDER CONDITIONS OF REPEATED TESTING

JAMES H. CROUSE[1]

University of Delaware

Two experiments investigated the acquisition of course material under conditions of repeated testing. In Experiment 1, with limited study intervals, acquisition increased over trials of study followed by testing. In addition, adjunct information about the content of the test item pool also increased performance. In Experiment 2, with student-determined study intervals, there was little or no change in performance associated with repeated testing.

The present educational system requires that students learn from written instructional materials. The common practice of introducing an assigned instructional passage in a course is a one-shot opportunity in which a single study interval is followed by a test, the results of which, hopefully, demonstrate that some learning has occurred. Under conditions of repeated testing, at minimum, a second study interval is introduced and is followed by a test; that is, a second trial is introduced which allows additional learning to occur and to be demonstrated. Repeated testing has been introduced by Jensen and Prosser (1969), but these investigators did not report any acquisition data. Acquisition data could be obtained with any of the computer-based instructional management systems that are currently in use (cf. Baker, 1971). Another possibility for obtaining trial-by-trial acquisition data would be to utilize the "mastery learning" strategy described by Bloom (cf. Bloom, Hastings, & Madaus, 1971) or the "personalized system of instruction" described by Keller (1968). In the present studies, an instructional passage was assigned to students, and for this passage, a large pool of test items was constructed to measure learning. A large number of tests were then constructed, each of which consisted of a random sample of test items from the pool, and each student was given a different one of these tests during acquisition on each trial.

[1] Requests for reprints should be sent to James H. Crouse, Department of Educational Foundations, University of Delaware, Newark, Delaware 19711.

Experiment 1

The main purpose of this experiment was to study acquisition under conditions of repeated testing. An instructional passage was available for study only in a study room. On each trial, the students would come to the room, study the passage for a two-hour period, and then take one of the tests on the material. In this way, the available study interval on each trial was fixed at two hours. The amount of actual study time cannot be determined, of course, but it is probably less than the available study interval. This procedure was followed five times for a total of five acquisition trials according to a schedule which each student individually prearranged prior to the first trial. It was expected that as the number of study trials increased, the amount of learning would increase.

The second purpose of this study was to determine the effects on acquisition of the amount of information given to students about the content of the test item pool. All of the items in the pool were generated by applying a single rule to the instructional passage. Therefore, all items in the pool from which the tests were constructed could be said to represent the same item type (Anderson, 1972; Bormuth, 1970; Schlesinger & Weiser, 1970). In the no information

condition, no information adjunct to the instructional passage was given to students about the content of the item pool from which their tests would be constructed. In the rule information condition, students were given a statement of the rule that was used to generate the test items in the pool, and illustrations of the use of the rule were also given. In the question information condition, students were given the actual test questions which comprised the pool. These conditions represented the amount of information given about the content of the test item pool and were similar to procedures which provide adjunct encoding cues to increase learning (cf. Crouse & Idstein, 1972). Acquisition was expected to increase over the no information, rule information, and question information conditions, respectively, since with increased information about the content of the question pool, more information required by output which might not otherwise be encoded should be encoded into memory (cf. Crouse & Idstein, 1972).

Method

Materials. An article entitled "Reforms As Experiments" by Campbell (1969) was chosen from the reading materials of a course in educational research procedures. A 117-question test item pool was constructed by applying the following rule 117 times throughout the passage. A sentence was taken directly from the passage and expanded, if necessary, to include words obviously understood from the context of the passage but not included in the sentence. From this expanded sentence, a portion was identified, and a wh- question was derived to question this portion. For example, the sentence, "Many of the difficulties lie in the intransigencies of the research setting and in the presence of recurrent seductive pitfalls of interpretation," was expanded according to the context of the article to read, "Many of the difficulties of an experimental approach to social reform lie in the intransigencies of the research setting and in the presence of recurrent seductive pitfalls of interpretation." From this expanded sentence, the wh- question, "Where do many of the difficulties of an experimental approach to social reform lie?", was derived. From the pool of 117 test questions, a large number of 21-item tests were constructed. The tests were constructed by a computer program which randomly sampled 21 items for each test from the pool of 117 items. The tests had a completion format in which each question was typed with a space where its answer was to be supplied by the student.

Procedure. All students in the experiment were enrolled in the course in educational research procedures. When each student reached the time in the course to study the article entitled "Reforms As Experiments," he was given a general set of instructions which explained to him that he was being asked to follow a very specific set of procedures in the hope that rapid learning would occur. Explanation of the procedures was in terms of course improvement, and no mention was made of an experiment. All studying was done in a special study room, and each student individually scheduled five study periods in the room prior to his first appearance. When each student arrived, he was handed the article and given exactly two hours to study it. Immediately following this study period, he was given one of the tests and allowed unlimited time to complete it. This procedure was followed each time the student came to the study room, and students came to the study room five times during a two-week period. Thus, each student was allowed 5 two-hour acquisition trials for learning. No information was given students about their acquisition progress until after they had completed the fifth acquisition trial.

When each student arrived at the study room the first time, he was assigned to one of the three information conditions: no information, rule information, or question information. Students in the no information condition completed five acquisition trials in the above manner with no information adjunct to the passage being given about the content of the test item pool. Students in the rule information condition were given a written statement of the rule that was used to generate the test items in the pool. Included with the statement were examples of the use of the rule, but these were applied to different written materials than the students would be reading. The students were told that the questions to test their learning would be constructed using this rule, and they were allowed to refer to the rule information as much as they wished during learning. The students in the question information condition were given the entire pool of 117 questions from which their tests would be drawn. The questions were listed in the order in which their answers occurred during reading, and students were allowed to refer to these questions freely during learning.

Design. Five acquisition trials were given under each of the three information conditions: no information, rule information, and question information. Therefore, the design was a 5×3 factorial with repeated measures on the first factor. Fourteen students enrolled in the course in educational research procedures at the University of Delaware were randomly assigned to each of the three information conditions.

Results

The number of correct answers on each acquisition trial was computed for each student. An answer was counted correct if it

TABLE 1

MEAN NUMBER OF CORRECT RESPONSES FOR EACH INFORMATION CONDITION ON EACH TRIAL

Information	Trial				
	1	2	3	4	5
No information	5.82	8.68	8.75	9.82	12.04
Rule information	4.65	7.18	9.18	8.54	10.65
Question information	8.11	12.00	13.93	16.04	16.90

was a verbatim reproduction or a complete and accurate paraphrase of the answer. It was counted one half correct if it failed to meet these criteria but was partially correct. Otherwise, it was counted as incorrect. The mean number of correct answers on each of the five acquisition trials is shown in Table 1 for each of the three information conditions. It can be seen that acquisition increased over the five trials for all conditions ($F = 55.42$, $df = 4/156$, $p < .01$). Also, acquisition differed significantly in the three information conditions ($F = 14.74$, $df = 2/39$, $p < .01$), and Newman-Keuls tests showed that performance was greater in the question information condition than in the rule information and no information conditions (both, $p < .01$) and that these latter conditions did not differ greatly ($p > .05$). Finally, the differences among the three information conditions increased over trials as indicated by their significant interaction with trials ($F = 2.66$, $df = 8/156$, $p < .01$).

Each student was given a random sample of test items from the question pool on each trial; therefore, there inevitably was some repetition of questions across trials. A question which arises is the extent to which the acquisition data in Table 1 might represent gains in performance associated with the repeated questions over trials. For example, students might remember an item from one trial when it is presented on a later trial, and acquisition might predominantly reflect gains of this kind. Consequently, acquisition was examined over Trials 2–5 separately for those items that were repeated from an earlier trial and those items that were not repeated. The number of items correct of each type was expressed as a percentage of the number of items which occurred of each type, and these percentages are summarized for each acquisition trial and information condition in Table 2. The results clearly indicate that acquisition increased over trials ($F = 10.79$, $df = 3/117$, $p < .01$). They also indicate that some item learning did occur since overall acquisition was higher for the repeated than nonrepeated items ($F = 18.86$, $df = 1/39$, $p < .01$); however, acquisition was not limited only to the repeated items, as indicated by the nonsignificant Trials × Repeated versus Nonrepeated Items interaction ($F < 1.00$). The amount of item learning interacted with the information conditions; item learning was greatest in the no information condition, somewhat less in the rule information condition, and least in the question information condition ($F = 3.87$, $df = 2/39$, $p < .03$). This latter result is to be expected since in the question information condition, all items were avail-

TABLE 2

PERCENTAGE CORRECT ANSWERS FOR NONREPEATED AND REPEATED ITEMS

Information	Nonrepeated items				Repeated items			
	Trial 2	Trial 3	Trial 4	Trial 5	Trial 2	Trial 3	Trial 4	Trial 5
No information	38.72	35.44	40.40	46.00	49.38	48.66	58.70	62.64
Rule information	31.76	42.06	32.53	52.36	41.56	41.76	49.84	51.73
Question information	58.59	63.35	75.09	78.65	64.41	69.46	71.07	80.75

able for inspection at all times during study. Again, it can be seen that overall acquisition differed among the three information conditions ($F = 14.36$, $df = 2/39$, $p < .01$).

Discussion

The results indicate that an increased number of acquisition trials resulted in increased learning of course material. They also indicate that increased information about the content of the test item pool facilitated performance. Almost all of this increase occurred when students were given the actual items from which their tests would be constructed (question information condition), with no apparent effect being seen when a rule was given by which the test items would be constructed (rule information condition). There are several possibilities to account for this finding: the rule may not have been helpful because the extreme extension of its use generated a potential test question for every sentence in the passage; efforts to use the rule might have been abandoned because they disturbed the students usual study habits; the rule may have been too complicated and was simply ignored; or other explanations are possible as well. Nevertheless, information about the specific items in the test pool did facilitate acquisition and did add further support to the principle that under circumstances where conditions adjunct to reading materials can plausibly be assumed to influence encoding of information required by output that would not otherwise be encoded, output performance will be facilitated (cf. Crouse & Idstein, 1972). The important element lacking in this conceptualization as well as in all studies investigating the effects of adjunct encoding cues is how to conceptualize the memory information that is required for performance on the different kinds of output questions used to measure learning from prose instruction. In the absence of this knowledge, the search for adjunct encoding cues that will facilitate encoding of relevant memory information and, therefore, increase performance on different item types will necessarily continue on a hit-or-miss basis.

Experiment 2

The results of Experiment 1 indicated that as the number of study trials increased, acquisition increased. Another way to say this is that students were motivated to do more learning than they could achieve on the early trials. This implies that the beneficial effects of repeated testing resulted from the use of a study interval as short as two hours on each trial, which restricted the amount of learning that could occur. Experiment 2 tested this possibility. Students were given the reading material so that they could take it home to study at their convenience. When they wanted to, they could come to the study room and be tested on the reading. In this setting, the available study interval was student determined by the length of time before they came in to be tested. The amount of actual study time which occurred during this period was estimated by having each student keep a detailed record of his study. After being tested, the students could again take the passage for a second study interval, study as much as they wanted, and then come for testing. A total of four acquisition trials could be completed in this fashion; the number each student completed was left optional.

Method

Materials. The materials were the same ones used in Experiment 1; that is, the article entitled "Reforms As Experiments," the 117-question test pool, and the batch of 21-question tests each comprising a random sample of 21 questions from the question pool.

Procedure. All students in the experiment were again enrolled in the course in educational research procedures. When each student reached the time in the course to study the article entitled "Reforms As Experiments," he was given general instructions similar to those in Experiment 1 which explained that he was being asked to follow very specific procedures and that these procedures were justified in terms of facilitating learning and effecting course improvement. No mention was ever made of an experiment. Studying could be done anywhere, and each student was asked to record, to the minute, each starting and stopping time. How to do this was explained as was the use of a form for recording. Short breaks from study were not to be included as study time. When ready, the students could come for testing, at which time they were given one of the tests with unlimited time for its completion. No information was ever

provided the student about the content of the test item pool. After completing a test, they could undertake additional study and follow it with a second test. Up to four trials could be completed in this fashion. Fifty-three students participated in this study-learning procedure.

Results

As expected, most students ($n = 36$) completed all four acquisition trials; the number of students who completed only one, two, or three trials was 5, 6, and 6, respectively. Because the number of students was substantial only for the four-trial group, subsequent analyses were done on these data. The number of correct answers on each of the four acquisition trials was computed for each student. Answers were scored in the same way as Experiment 1; that is, fully correct for a verbatim reproduction or completely accurate paraphrase, one half correct for a failure to meet the fully correct criteria but a partially correct answer, and incorrect. The mean number of correct answers was 4.26, 5.41, 5.30, and 4.98 on Trials 1–4, respectively. These data show that performance first increased slightly and then decreased slightly over trials. While the analysis of variance revealed a significant F for these changes ($F = 9.27$, $df = 3/105$, $p < .01$), the overall magnitude of these changes was very small; that is, always less than 1.5 correct answers out of a possible 21. It is as if the potentially unlimited study intervals, in contrast to Experiment 1, allowed the students to acquire as much learning as they wished on Trial 1 but that with added trials they did not acquire more.

The question arises as to the relationship between study time and acquisition. The actual time spent studying was available for each student, and since the greatest variability in study time was on the first trial, this first-trial data seemed most likely to reveal the expected positive relationship between the length of study time and performance. Unexpectedly, there was no sign of a positive correlation ($r = .10$, $df = 34$, $p > .05$). One explanation for this failure to find a positive relationship between study time and acquisition is that study time is negatively correlated with ability and that ability is, in turn, directly related to correct performance. Consequently, the positive relationship between study time and performance could be masked by ability differences. The data supported this interpretation: for the 28 students completing four acquisition trials for whom measures of ability were available (verbal scores on the Graduate Record Examinations), the correlation between study time on Trial 1 and ability was in the right direction and was significant ($r = -.50$, $p < .01$). The correlation between ability and correct performance on Trial 1 was also in the right direction but fell just short of significance ($r = .30$, $p < .10$). Most importantly, however, when differences in ability were partialed out, the correlation between study time and correct performance was positive and significant ($r = .42$, $p < .01$).

The results of both experiments lend further support to the importance of study time as a factor in learning. In Experiment 1, with limited study intervals, study time increased over trials as did acquisition. In Experiment 2, with potentially unlimited study intervals, there was little increase beyond Trial 1; but on Trial 1, study time was found to be positively related to performance after differences in student ability were partialed out. It seems that repeated testing as used in its barest form in the present studies is important primarily as a way of increasing study time when the study intervals are less than optimal for achieving the desired learning. If study time can be increased by other means, little or no advantage may be found for repeated testing, at least as used in the present studies.

REFERENCES

ANDERSON, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research*, 1972, **42**, 145–170.

BAKER, F. B. Computer-based instructional management systems: A first look. *Review of Educational Research*, 1971, **41**, 51–70.

BLOOM, B. S., HASTINGS, J. T., & MADAUS, G. F. *Handbook on the formative and summative evaluation of student learning*. New York: McGraw-Hill, 1971.

BORMUTH, J. *On the theory of achievement test items*. Chicago: University of Chicago Press, 1970.

CAMPBELL, D. T. Reforms as experiments. *American Psychologist*, 1969, **24**, 409–429.

CROUSE, J., & IDSTEIN, P. Effects of encoding cues

on prose learning. *Journal of Educational Psychology*, 1972, **63**, 309–313.

JENSEN, D., & PROSSER, F. Computer-generated, repeatable examinations and large class instruction. Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago, May 1969.

KELLER, F. S. "Goodbye teacher..." *Journal of Applied Behavioral Analysis*, 1968, **1**, 79–89.

SCHLESINGER, I. M., & WEISER, Z. A facet design for tests of reading comprehension. *Reading Research Quarterly*, 1970, **5**, 566–580.

(Received July 13, 1973)

Manuscripts Accepted for Publication in the *Journal of Educational Psychology*

Effects of the Availability of Objectives and/or Rules on the Learning Process. Paul F. Merrill (Center for Educational Design—Computer Support of Instruction, Room 10 Kellum Hall, Florida State University, Tallahassee, Florida 32306).

Relationship of Three Cognitive Style Dimensions to Elementary Reading Abilities. Douglas R. Denney (Department of Psychology, University of Kansas, Lawrence, Kansas 66045).

Signal Systems of Lesson Settings and the Task-Related Behavior of Preschool Children. Jacob S. Kounin (Department of Educational and Clinical Psychology, Wayne State University, Detroit, Michigan 48202) and Paul V. Gump.

"Environments" of British Commonwealth Universities. James M. Richards, Jr. (Center for Study of Social Organization of Schools, The Johns Hopkins University, Baltimore, Maryland 21218).

Multidimensional Scaling of Concept Learning in an Introductory Course. Howard Wainer (5848 South University Avenue, Chicago, Illinois 60637) and Kenneth Kaye.

Some Problems in the Application of Achievement Motivation to Education: The Assessment of Motive to Succeed and Probability of Success. Charles B. Schultz (Department of Education, Trinity College, Hartford, Connecticut 06106) and Michael Pomerantz.

Course Evaluation: When? Neil A. Carrier (Department of Psychology, Southern Illinois University, Carbondale, Illinois 62901), George S. Howard, and William G. Miller.

Impact of Coeducation on "Fear of Success" Imagery Expressed by Male and Female High School Students. Ronny Winchel, Diane Fenner, and Phillip Shaver (Department of Psychology, Columbia University, Schermerhorn Hall, New York, New York 10027).

Self-Determination of Academic Standards by Children: Toward Freedom from External Control. Jeffrey J. Felixbrod (Department of Psychology, State University of New York, Stony Brook, New York 11790) and K. Daniel O'Leary.

Training for Thematic-Fantasy Play in Culturally Disadvantaged Children: Preliminary Results. Eli Saltz (Center for the Study of Cognitive Processes, Wayne State University, Detroit, Michigan 48202) and James Johnson.

Relationships Among Perceptual-Motor Tasks: Tracing and Copying. Marcia Bernbaum, Jacqueline Goodnow, and Elyse Lehman (Department of Psychology, George Washington University, 2125 G Street, N.W. Washington, D.C. 20006).

Another View of the Relation of Environment to Mental Abilities. Chester W. Harris (Department of Education, University of California, Santa Barbara, California 93106) and David L. McArthur.

Another View of the Relation of Environment to Mental Abilities: A Reply. Kevin Marjoribanks (Department of Educational Studies, University of Oxford, 15 Norham Gardens, Oxford, England OX2 6py).

Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Donald B. Rubin (Division of Data Analysis Research, Educational Testing Service, Princeton, New Jersey 08540).

Differential Effects of Reward versus No Reward Instructions on the Creative Thinking of Two Economic Levels of Elementary School Children. Roger A. Johnson, School of Education, Old Dominion University, Norfolk, Virginia 23508).

The Influence of Mode of Presentation, Ethnicity, and Social Class on Teachers' Evaluations of Students. Mary Jensen (School of Physical Education, Dalhousie University, Halifax, Nova Scotia, Canada) and Lawrence B. Rosenfeld.

Journal of Educational Psychology
1974, Vol. 66, No. 3, 373-385

PREDICTABILITY OF STUDENTS' EVALUATIONS OF COLLEGE TEACHERS FROM COMPONENT RATINGS[1]

GRACE FRENCH-LAZOVIK[2]
University of Pittsburgh

Two similarly designed studies which were conducted 15 years apart at different universities and which involved over 9,700 students and 277 faculty gave nearly identical answers to the question of what teaching characteristics carry greatest weight in predicting students' general opinion of their teachers. Items used on student evaluation of teaching scales were treated as predictors of students' overall ratings of teaching effectiveness. Reduced-rank regression analysis revealed high multiple correlations (.97 and .93) for items dealing with clarity of exposition, arousal of student interest, and stimulation or motivation to intellectual activity. Neatness of appearance, friendliness of manner, sense of humor, the giving of individual attention, and the handling of examinations carried little weight in predicting students' evaluations of effective teaching.

The use of student evaluations of college teaching, either systematically or informally, as measuring devices has spread rapidly on college campuses in recent years. Doubtless, pressures for accountability and student unrest have both contributed to this increase. Whatever purpose these evaluations are intended to serve—as feedback to teachers for their own improvement or as information to department chairmen for decisions—a better understanding of what students are evaluating in teaching is sorely needed.

Much is known about the technical characteristics of student evaluations, for example, their reliability, correlations with class and student variables, and validity defined as correlation with measures of performance (for excellent reviews of this literature, see Costin, Greenough, & Menges, 1971; McKeachie, 1969). However, when human judgment is used as a measuring device, there are questions bearing on construct validity which should be asked. Could this complex judgment be analyzed into the important components which contribute to it? What is the relative influence of each component? In the case of student judgments, what kind of teaching do students evaluate most highly?

Many investigators have approached this last question by asking a similar though different one; that is, what do college students consider to be the most important characteristics of good teaching? The late Edwin R. Guthrie (personal communication, approximate date, June 1947), whose unpublished studies of student evaluations of teaching began in the 1920s,[3] sought answers to this question by simply asking students to describe the qualities they considered of greatest importance in good teaching. A first list was condensed by content analysis and presented to a second group of students who were asked to check those

[1] The University of Washington study was supported in part by a grant from the Carnegie Corporation to Edwin R. Guthrie and in part by Public Health Research Grant M-743(C2) and Office of Naval Research Grant Nonr-477(08) to Paul Horst. The University of Pittsburgh study was supported by the Center for the Improvement of Teaching, College of Arts and Sciences.

[2] Requests for reprints should be sent to Grace French-Lazovik, Center for Improvement of Teaching, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

[3] Some of these early studies resulted in the five-item scale used at the University of Washington from 1946 through 1956. This scale appears in the well-known monograph by Guthrie (1954).

qualities which were characteristic of their best teacher.

This methodology, or minor variations of it, has been used by many subsequent investigators (Crawford & Bradshaw, 1968; Downie, 1952; Gadzella, 1968; Musella & Rusch, 1968; Riley, Ryan & Lifshitz, 1950; Smith, 1944) despite the fact that it leaves much to be desired, since it relies on the students' ability to answer the question adequately, provides no way of knowing whether all relevant variables have been included, and does not reveal the relative differences in importance of the characteristics mentioned or checked.

Factor-analytic studies of item pools in this domain typically do not include a criterion measure. Thus, they give information on the dimensions being measured but not on which of these dimensions relate to the students' concept of good teaching. It is quite likely that some of the dimensions measured do not contribute to the discrimination made by students between good and poor teachers.

In an early classic study, Coffman (1954) avoided both the problem of having students introspect as to what they valued in teaching and that of factor validity by factor analyzing the 18-item *Oklahoma A. and M. Rating Scale for Instructors* to determine what factors accounted for the students' general estimate of teaching effectiveness. While this study provided a better methodology, there are two aspects of it in which major improvements can be made. First, Coffman's items were relatively few in number and were limited to those on one existing scale. Second, a more powerful design would result by simply posing the same question in terms of a prediction model. The students' general estimate of teaching effectiveness could be examined to determine how well it can be predicted, what the best predictors are, and what the relative weight of each item in the prediction is.

Such a formulation involves several problems. In the study of student evaluations, approximately 40 to 50 predictor variables would have to be considered and data on these variables could feasibly be obtained for perhaps 100 to 150 teachers. The problem of degrees of freedom then arises, for whenever the number of variables is large in comparison to the number of cases, the determination of regression weights is usually considered not justified, since a stable solution cannot be expected. Fortunately, this problem can be avoided by using the reduced-rank regression model developed by Horst (1941). In his procedures, regression coefficients are based on the principal-components factor matrix which usually has a considerably smaller rank than the intercorrelation matrix. In addition, Leiman (1951) and Burket (1964) have shown that reduced-rank regression analysis results in more stable regression weights than those based on traditional procedures. A special advantage for the study of judgments derives from the fact that regression coefficients are calculated from an orthogonal matrix. As a result, the product of the regression coefficient and of an item's correlation with overall judgment can be rank ordered to indicate the relative contribution of each item to the overall prediction.

The recent University of California, Davis, study (Hildebrand, Wilson, & Dienst, 1971) did use a large item pool. However, the efficacy of their factor-analytic findings rests in large part on their method of item selection, namely, the use of discrimination indices. In addition to the fact that the stringency of their criteria for inclusion in the extreme groups can be questioned, these indices pose problems related to sample composition (Gulliksen, 1950).

Further, Hildebrand et al. (1971) did not use methods which indicate the relative importance of the different criteria students use in evaluating teaching.

The application of reduced-rank regression analysis to a broadly representative item pool should further our understanding of what kind of teaching is evaluated most highly by students.

Problem

The two nearly identical studies reported herein were designed to determine what teaching characteristics are most predictive of college students' overall judgments of teaching effectiveness. The first study was

carried out at the University of Washington, in the 1956–1957 academic year, the second at the University of Pittsburgh in the 1971–1972 academic period.

Method

For this problem, the application of the reduced-rank regression model required that students register judgments of their teachers on a set of verbal descriptions phrased as items on a rating scale and that they also register judgments of their instructor's overall teaching effectiveness. The separate items were then treated as a set of predictors of the overall judgment. By factor analyzing the predictor intercorrelations, using the principal-axis factor matrix to calculate the regression coefficient for each item, and obtaining the product of these values and their respective correlations with the criterion, the relative contribution of each item in predicting the overall judgment was determined.

Statements describing teacher characteristics or classroom behavior were phrased so that they could be judged on a 5-point scale. The descriptive categories were stated in comparative terms that were the same for all items and for the overall judgment. The categories used were as follows:

In this respect, your instructor ranks below most of the teachers you have known.
In this respect, your instructor is only fair in comparison with other teachers you have known.
In this respect, your instructor is competent and compares well with the average of the teachers you have known.
In this respect, your instructor is well above the average of the teachers you have known.
In this respect, your instructor is one of the most outstanding you have known.

It was important that the original set of items span the field of variables which could possibly be related to students' concepts of good teaching. The initial pool depended on the investigator's guesses as to what the important variables might be, but the method of analysis provided a precise test of the degree to which the chosen items did, in fact, include all relevant variables; that is, the analysis indicated what percentage of the variance of the overall judgments was accounted for by the set of items.

Item Collection: University of Washington Study

To increase the probability that all relevant variables would be tapped, items were sought from different sources. The top 10 items from Guthrie's studies were included (see Footnote 3). A representative set of items was chosen from the rating forms used at many colleges and universities throughout the country. Students' general comments about their teachers, collected over the years from open-ended questions on the teaching evaluation form used at the University of Washington, were also searched. And lastly, items based on faculty members' views regarding important characteristics of good teaching were added. A final set of 41 verbal statements represented the end product of the collecting, modifying, and rephrasing procedures (see Table 1).

Item Collection: University of Pittsburgh Study

Instead of sampling sources of items, the University of Pittsburgh study tried to represent some of the factors found in factor-analytic studies of student evaluation item pools (Bendig, 1954; Coffman, 1954; French, 1957; Hildebrand et al., 1971; Hodgson, 1958; Isaacson et al., 1964; Rehberg, Roberts, & Vandament, 1970; Remmers & Baker, 1952; Solomon, 1966). While the number of factors identified by these investigators ranged from 3 to 10, 4 to 6 of them appeared rather consistently. Authors showed little congruence in the names assigned to the various dimensions, so for the purposes of this study, 6 item clusters were chosen and renamed. These clusters were identified as follows: (A) arousal of student interest, (B) clarity, (C) encouragement of student initiative, (D) teacher preparation, (E) teacher interest in students, and (F) grading. Items representing these clusters are marked by the appropriate letter in Table 1. The following 5 items, not used in the University of Washington study, were also represented in the clusters:

Has stimulated thinking on the part of the students (C),
Has presented course content in an organized manner (D),
Bases grades on adequate sample of student work (F),
Gives feedback as to how students are progressing (F), and
Has presented course content so that you perceive its relevance to your interests (new item).

The 16 items in Table 2 were selected as a subset to represent these 6 clusters.

University of Washington Sample

A 25% sample of the faculty, exclusive of graduate professional schools, was selected so as to be representative of each subject taught and of each of the academic ranks from instructor through full professor. Letters requesting cooperation in the study and providing assurance that the results would be used only for research were written to the 144 faculty members selected; 133 or 92% expressed willingness to participate. In the classes of 37 professors, 22 associate professors, 52

assistant professors, and 22 instructors, 3,654 rating forms were completed and provided the data base of the University of Washington study.

University of Pittsburgh Sample

A random sample consisting of 25% of each of the academic ranks from instructor through full professor in the College of Arts and Sciences was selected. These 160 faculty were asked permission to survey one of their undergraduate classes. Of these, 144 or 90% were surveyed and 6,120 rating sheets were completed in the classes of 42 professors, 32 associate professors, 51 assistant professors, and 19 instructors.

Data Collection

In both studies, surveys of student opinion of teaching were arranged at a time within the last three weeks of the term which was convenient to each teacher. Precautions were taken to guarantee the student's anonymity, to hold constant the instructions to each class, and to dissociate the surveys from any administrative purpose in the minds of both the students and the faculty. Administering the questionnaire entailed going into the classroom of each instructor at the appointed time, waiting for him or her to leave, giving explanations about the survey, distributing forms, answering questions, and receiving the completed form from each student present.

Analyses

Data from the two studies were analyzed separately following identical designs. Analysis began with the computation of the average, for each item, of the ratings assigned by the members of a class to their teacher. These averages were intercorrelated for the predictor items, and this matrix was factor analyzed using Hotelling's (1933) principal-components method. Factoring was stopped when the sum of the eigenvalues was approximately equal to the sum of the item reliabilities, calculated by Horst's (1949) generalized formula. Regression coefficients were then obtained for the predictors using the following formula (Horst, 1941):

$$B = a\, D^{-2}\, (a'\, r_c)$$

where

- B = the vector of regression coefficients,
- a = the principal-components factor-loading matrix,
- a' = the transpose of a,
- D = the diagonal matrix which is the product of $a'a$, and
- r_c = the vector of criterion correlations.

The elements of the Vector B are designated B_i and of the Vector r_c, r_{ic}. The quantity $B_i r_{ic}$ represents the product of the regression coefficient and the validity coefficient for each item, and these values were rank ordered. Lastly, the multiple correlation of the predictors with the overall judgment was determined for each study.

Results

Table 1 presents the findings at the University of Washington, and Table 2 gives the results at the University of Pittsburgh.

It is not surprising that the number of factors differs in the two studies since the item-sampling procedures were not the same. The high multiple correlations shown in Tables 1 and 2 (R = .944 and .976, respectively) established that student judgments of overall teaching effectiveness can be predicted with very considerable accuracy.

In answering the question, "Were all relevant variables included in the set of items chosen?", it can be seen that since the two multiple correlations are very close to the maximum obtainable, all of the reliable variance was accounted for. Thus, no variable of importance in predicting student overall judgment of teaching effectiveness was left unmeasured in either study. Other sets of items, differently worded, might predict overall teaching effectiveness equally well, but they would not include variables different from those measured here.

With this established, the major problem of these studies, that is, the relative importance of the predictors, can be examined by inspection of the rank ordering of items in Tables 1 and 2. (The number in parentheses following each item indicates rank order.) Comparison of the high-ranking items on the two lists gives a picture of surprising consistency across a 15-year time span on two quite different campuses with different student and faculty populations. Even more interesting is the fact that 5 of the items taken from Guthrie's (1954) list for the University of Washington study (knowledge of subject, arousal of student interest, teacher's interest in subject, good use of examples and illustrations, clarity of explanations) are in the top 10 of Table 1, and the 4 items from his study used in the Uni-

versity of Pittsburgh study (clarity, arousal of student interest, use of examples, knowledge of subject) are the top 4 of Table 2. These investigations span a period of over 40 years and represent different generations of students. It is apparent that some of the criteria used by college students in judging their teachers are surprisingly resistant to change.

Four items on the University of Washington list suggested by faculty (new viewpoints or appreciations, broadening of interest, increased skills in thinking, new tools for attacking problems) ranked 1, 3, 8, and 12, respectively. The University of Pittsburgh study used only two of these items (new viewpoints or appreciations, new tools) and they ranked 6 and 13. There is, then, evidence of some degree of agreement between students and faculty as to what constitutes good teaching.

The content of the 10 top-ranking items on the two lists is nearly identical. Every one of the top 10 items on the University of Pittsburgh list is either a duplicate, a modification, or from the same cluster (Table 1) as the top 10 of the University of Washington study. These verbal statements relate to clarity, organization, teacher's knowledge, use of examples, arousal or broadening of interests, and motivation or stimulation of thought.

It is informative to note the position in Table 1 of statements dealing with the handling of examinations. In both studies, these items carry little or no weight in predicting teaching effectiveness. Nor is this criterion predicted by such qualities as neatness of appearance, friendliness, sympathetic manner, sense of humor, or the giving of individual attention. Items like these appear on many scales that are used to ascertain student opinion of teaching. The evidence here suggests their presence can mislead instructors into believing that improvements of these characteristics would be followed by higher general evaluations.

Discussion

Before interpreting the results, it is important to ask, could the high multiple correlations obtained simply reflect halo effect? The tendency of raters to respond similarly to all items on the basis of a global impression results in high item intercorrelations that reflect the presence of a general factor. Hodgson (1958) applied rotational models to the University of Washington factor-loading matrix and found that a solution involving one general factor and eight primary factors gave the best approximation to simple structure. The general factor accounted for approximately 5% of the variance, and Hodgson postulated that it represented halo effects. Thus, if halo or response style variables account for the general factor, their effect is so small that it cannot be regarded as the basis of the high multiple correlations obtained. (Additional evidence on this point is presented in the Discussion section.)

Interpretation of the University of Washington and University of Pittsburgh results can be facilitated by focusing, in both studies, on the content of the items that are most predictive of students' overall judgments of teaching effectiveness. The content of the top 10 items on the two lists can be structured into larger categories. They could, for instance, be considered under the following three categories: clarity of exposition, arousal or broadening of interest, and motivation of intellectual activities or stimulation of thought.

Clarity of exposition:
- Interprets difficult or abstract ideas clearly
- Makes good use of examples and illustrations
- Has presented the course in an organized manner
- Inspires confidence in his/her knowledge of the subject

Arousal of student interest:
- Increased my interest in the subject
- Gave me new viewpoints or appreciations
- Broadened my interests
- Included worthwhile material not duplicated in the text
- Has presented many thought-provoking ideas
- Shows interest and enthusiasm in his subject

TABLE 1

FACTOR-LOADING MATRIX AND ITEM RELIABILITIES, VALIDITIES, AND PREDICTIVE WEIGHTS FOR THE UNIVERSITY OF WASHINGTON STUDY

Item	Cluster[a]	Factor[b]								Item characteristics			
		I	II	III	IV	V	VI	VII	VIII	r_{ii}	r_{ic}	B_i	$B_i r_{ic}$
1. Has an adequate speaking voice (30)[c]		.567	−.154	−.083	.157	.219	−.526	−.011	.193	.912	.573	.019	.011
2. Shows respect for the questions and opinions of others (41)	E	.341	.306	−.170	−.311	.555	−.221	.079	.152	.876	.274	.005	.001
3. Is willing to give individual attention (38)	E	.207	.263	−.004	−.585	.060	−.252	.428	.364	.852	.193	.021	.004
4. Has a good sense of humor (33)		.577	.484	.184	.221	.179	−.010	−.217	.006	.902	.542	.016	.008
5. Is sensitive to class difficulties (28)	E	.734	.487	−.138	−.114	−.194	−.012	.062	−.067	.839	.671	.023	.015
6. Stresses important material (18)	B	.771	−.173	−.201	−.157	−.249	.023	.051	.051	.801	.717	.031	.022
7. Has been well prepared for each day's lecture (15)	D	.664	−.562	−.144	−.085	.088	.236	.055	−.022	.830	.647	.037	.024
8. Avoids sarcasm or criticism which embarrasses a student in the presence of others (40)		.455	.400	−.354	−.179	.304	.224	.084	−.417	.909	.393	.005	.002
9. Takes an active, personal interest in the progress of his class (22)	E	.747	.369	.041	−.217	−.190	−.016	.022	−.007	.864	.689	.026	.018
10. Has given me new tools for attacking problems (12)	C	.821	−.032	.283	−.092	−.233	−.181	.061	−.077	.825	.810	.040	.032
11. Encourages initiative on the part of the students (14)	C	.779	.096	.399	−.128	−.171	−.058	−.013	−.040	.816	.729	.038	.028
12. Keeps the balance between discussion and lecture well suited to this type of course (20)	D	.839	.095	−.122	−.040	−.115	−.046	−.010	−.054	.824	.760	.029	.022
13. Has made the course sufficiently difficult to be stimulating (21)	C	.760	−.292	.250	−.274	.083	−.001	−.093	−.197	.812	.740	.029	.021
14. Is reasonable and fair in handling of examinations (35)	F	.669	.300	−.208	−.212	−.155	.270	−.254	.306	.721	.597	.012	.007
15. Avoids unnecessary repetition (36)	D	.729	−.221	−.038	−.314	.136	−.069	−.286	−.028	.808	.660	.009	.006
16. Uses class time efficiently (23)	D	.736	−.458	−.155	−.269	.087	.017	−.031	.057	.847	.662	.027	.018

17. Makes challenging assignments (34)	C	.606	−.233	.442	−.371	.007	−.171	−.201	−.211	.860	.570	.015	.008
18. Seems to enjoy teaching (17)		.750	.130	.239	.147	.247	−.080	−.124	.072	.828	.710	.032	.023
19. Is fair in grading (37)	F	.650	.305	−.152	−.219	−.138	.290	−.331	.305	.807	.586	.009	.005
20. Is friendly and sympathetic in manner (31)	E	.650	.598	−.156	.011	.158	.076	.013	−.196	.852	.574	.017	.010
21. Teaches near the class level (19)		.777	.207	−.292	.146	−.269	−.123	.140	−.042	.824	.703	.031	.022
22. Is free from annoying mannerisms (32)		.755	.061	−.280	.214	.110	−.232	−.155	−.081	.856	.673	.014	.009
23. Gives well organized lectures (24)	D	.738	−.467	−.327	−.091	.096	.038	−.032	−.039	.868	.707	.025	.018
24. Interprets abstract ideas and theories clearly (6)	B	.886	−.176	−.138	.116	.033	.011	.098	.044	.820	.856	.043	.037
25. Motivated me to do my best work (10)	C	.901	.114	.151	−.086	−.188	−.019	.027	−.092	.792	.831	.039	.033
26. Has increased my skills in thinking (8)	C	.864	.008	.293	−.104	−.191	−.079	.021	−.092	.809	.841	.041	.034
27. Makes the course practical (16)		.778	−.026	−.100	.140	−.277	−.183	.027	.068	.836	.719	.033	.024
28. Controls class discussion to prevent rambling and confusion (26)	D	.745	−.439	−.127	−.081	−.017	−.066	−.142	.032	.791	.683	.023	.016
29. Has given me new viewpoints or appreciations (1)	A	.872	−.046	.201	.147	.043	.107	.221	−.033	.824	.838	.060	.050
30. Is clear and understandable in his explanations (13)	B	.849	−.024	−.301	.183	−.047	−.087	.105	.006	.873	.799	.037	.029
31. Inspires class confidence in his knowledge of the subject (4)	B	.866	−.218	.044	.052	.140	.200	.070	.061	.876	.856	.050	.043
32. Is neat in appearance (39)		.564	−.092	−.231	.316	.019	−.364	−.176	−.020	.893	.483	.008	.004
33. Gets students interested in his subject (2)	A	.913	.035	.175	.153	−.009	−.014	.187	.000	.842	.880	.056	.050
34. Makes good use of examples and illustrations (5)	B	.781	−.116	.004	.228	.047	.159	.203	.111	.781	.754	.055	.042
35. Has presented many thought-provoking ideas (7)	A	.754	−.017	.452	.064	.136	.109	.034	−.036	.846	.743	.049	.036
36. Has maintained an atmosphere of good feeling in the class (27)	E	.757	.530	−.098	.068	.098	.041	−.034	−.125	.890	.716	.022	.016
37. Relates the subject matter to other fields (25)	A	.643	.151	.158	.393	−.086	.004	−.194	.187	.767	.615	.029	.018
38. Shows interest and enthusiasm in his subject (9)	A	.732	−.029	.354	.135	.316	.167	.007	.142	.838	.700	.049	.034

Table 1—*Continued*

Item	Cluster[a]	Factor[b]								Item characteristics			
		I	II	III	IV	V	VI	VII	VIII	r_{ii}	r_{ic}	B_i	$B_i r_{ic}$
39. Lectures permit good note taking (29)	D	.601	−.352	−.415	.021	−.058	.103	.015	−.120	.828	.579	.020	.012
40. Has helped broaden my interests (3)	A	.863	−.081	.087	.189	−.008	.145	.207	−.015	.771	.812	.058	.047
41. Includes in his lectures worthwhile and informative material which is not duplicated in the text (11)	A	.690	−.149	.044	.158	.094	.258	.070	.249	.827	.688	.048	.033

Note. The reliability of the criterion, overall teaching effectiveness, was .892. The $\Sigma (B_i r_{ic}) = .891$ and $R = \sqrt{\Sigma} = .944$. Abbreviations: B = the vector of regression coefficients; r_c = the vector of criterion correlations; the elements of the Vector B are designated B_i and of the Vector r_c, r_{ic}; the quantity $B_i r_{ic}$ represents the product of the regression coefficient and the validity coefficient for each item, and these values were rank ordered.

[a] Letters indicate item clusters sampled in the University of Pittsburgh study. These clusters were identified as follows: (A) arousal of student interest, (B) clarity, (C) encouragement of student initiative, (D) teacher preparation, (E) teacher interest in students, and (F) grading.

[b] The eigenvalues for Factors I–VIII are as follows: I = 21.896; II = 3.308; III = 2.148; IV = 1.769; V = 1.304; VI = 1.207; VII = .915; and VIII = .902.

[c] Number in parentheses indicates ranking of item on the basis of predictive weight ($B_i r_{ic}$). This ranking is slightly different from that reported by French (1957). The reason is that a complete reanalysis of the University of Washington data revealed a small error which produced a minor shift in the rank position of some of the items in the top ten. Two items (both dealing with clarity—clear and understandable in explanations, and stress of important material) moved downward with the consequent upward shift of two others (both dealing with the arousal of interest—shows interest and enthusiasm in his subject, presented thought-provoking ideas). Thus, the content of the items at the top is not altered. The best hypothesis as to the source of the error, in view of the reanalysis, appears to be that in transferring the original data to punched cards, errors were made in two columns by the punching service.

TABLE 2

FACTOR-LOADING MATRIX AND ITEM RELIABILITIES, VALIDITIES, AND PREDICTIVE WEIGHTS FOR THE UNIVERSITY OF PITTSBURGH STUDY

Item	Cluster[a]	Factor[b]					Item characteristics			
		I	II	III	IV	V	r_{ii}	r_{ic}	B_i	$B_i r_{ic}$
1. Has increased my interest in the subject (4)[c]	A	.938	−.169	−.172	−.019	−.026	.848	.933	.092	.0854
2. Has given me new viewpoints or appreciations (6)	A	.982	−.276	−.201	−.096	.022	.852	.868	.091	.0787
3. Includes in the lecture worthwhile and informative material which is not duplicated in the text (5)	A	.841	−.242	−.324	−.102	.035	.866	.828	.101	.0839
4. Interprets difficult or abstract ideas clearly (1)	B	.916	.117	−.151	−.073	.049	.843	.919	.104	.0959
5. Makes good use of examples and illustrations to clarify concepts (2)	B	.875	.004	−.221	−.090	.086	.823	.880	.106	.0931
6. Inspires class confidence in his/her knowledge of the subject (3)	B	.870	.076	−.128	−.155	.135	.838	.856	.103	.0876
7. Encourages initiative on the part of the students (14)	C	.834	−.242	.140	.390	−.099	.870	.764	.033	.0255
8. Has given me new tools for attacking problems (13)	C	.884	−.130	.096	.125	−.204	.840	.832	.048	.0399
9. Has stimulated thinking on the part of the students (10)	C	.906	−.253	−.085	.196	−.072	.811	.884	.069	.0608
10. Has presented course content in an organized manner (8)	D	.647	.680	−.200	−.006	−.118	.834	.682	.099	.0672
11. Uses class time effectively (7)	D	.710	.609	−.146	.117	−.038	.854	.727	.097	.0702
12. Shows respect for questions and opinions of students (11)	E	.811	.036	.275	.093	.446	.852	.758	.063	.0481
13. Is sensitive to student difficulties in understanding course material (12)	E	.844	.126	.353	.097	.232	.810	.800	.050	.0401
14. Is reasonable and fair in the handling of examinations and grading (16)	F	.663	−.023	.586	−.382	−.104	.877	.581	−.001	−.0003
15. Gives useful information as to how students are progressing (15)	F	.790	.166	.387	.033	−.256	.784	.717	.021	.0151
16. Has presented course content so that you perceive its relevance to your interests (9)		.873	−.167	−.059	−.175	−.136	.858	.841	.073	.0611

Note. The reliability of the criterion, overall teaching effectiveness, was .891. The $\Sigma\ (B_i r_{ic}) = .9531$ and $R = \sqrt{\Sigma} = .976$. Abbreviations: B = the vector of regression coefficients; r_c = the vector of criterion correlations; the elements of the Vector B are designated B_i and of the Vector r_c, r_{ic}; the quantity $B_i r_{ic}$ represents the product of the regression coefficient and the validity coefficient for each item, and these values were rank ordered.

[a] Letters indicate item clusters sampled in University of Pittsburgh study. These clusters were identified as follows: (A) arousal of student interest, (B) clarity, (C) encouragement of student initiative, (D) teacher preparation, (E) teacher interest in students, and (F) grading.

[b] The eigenvalues for Factors I–V are as follows: I = 11.160; II = 1.228; III = 1.058; IV = .473; and V = .450.

[c] Number in parentheses indicates ranking of item on the basis of predictive weight ($B_i r_{ic}$).

Presented course so that you perceive its relevance to your interests

Motivation or stimulation of thinking:
Stimulated thinking on the part of students
Increased my skills in thinking
Motivated me to do my best work

Some of the items grouped with clarity suggest means by which clarity can be achieved—organization of content, use of examples, conveying the teacher's knowledge. There are, of course, many other contributors to clear exposition.

It should be remembered that all of the students in the University of Pittsburgh study and most of those in the University of Washington study were undergraduates. At this level of academic maturity, students depend on their teachers for some degree of motivation in intellectual activities and for their involvement in the subject content. If these investigations were repeated on graduate students, who are presumed to be more self-motivated and to bring with them a developed interest in their field, the findings might be quite different. Similarly, other differences could be expected for evening and summer school students.

If the content categories have been correctly identified, then it should be possible, using only 1 item from each category, to approach the multiple correlation obtained with the full set of predictors. To test this possibility, a reanalysis was performed on the University of Washington data using the following items: (*a*) Interprets abstract ideas and theories clearly; (*b*) Gets students interested in the subject; and (*c*) Has increased my skills in thinking. The multiple correlation for these 3 items with the criterion is .928, while that for the set of 41 predictors is .944. In the University of Pittsburgh study, the 3 items used in a reanalysis were (*a*) Interprets difficult or abstract ideas clearly; (*b*) Has increased my interest in the subject; and (*c*) Has stimulated thinking on the part of students; for these 3 items, a multiple correlation of .965 resulted. This compares with a multiple correlation of .976 for the set of 16 predictors. Further evidence to support Hodgson's (1958) conclusion that halo effects were small can be offered here. If the results of halo accounted for these high multiple correlations, then any subset of 3 items should give equally high multiple correlations. The multiple correlation for 3 items from Table 1 whose regression coefficients show that they carry little or no weight in predicting the criterion (willing to give individual attention, avoids sarcasm, and respects students' opinions) is .421. This may be taken as an estimate of the maximum halo effect, though its true effect is probably less. Thus, the difference between .42 and .93 (or about 70% of the variance of the general estimate of teaching effectiveness) cannot be attributed to halo effect.

These analyses offer strong evidence that clarity, the arousal of interest, and the motivation to intellectual activity are the major determiners of students' general evaluation of teaching effectiveness. (It should be emphasized that the purpose of these analyses was to validate the categories, not to suggest that only three items should be used on a rating scale. Ideally, several items from each category should be used.)

The categories described are not meant to imply independence. On the contrary, they would seem to be complexly intertwined. Activities of the teacher which contribute to one of these objectives may simultaneously achieve one of the others. In using examples and illustrations to clarify a concept, the effective teacher chooses examples to add interest or to help students perceive relationships to their own experience. Clarification can also contribute to new viewpoints or appreciations. It is possible that this interdependence may account for the fact that items from all three categories often have their highest loadings on the large first factor found in so many of the factor-analytic studies of this domain.

The University of Washington and University of Pittsburgh results show congruencies with certain other findings. Where investigators asked what students valued in teaching, clarity or the qualities and activities which contribute to it (teacher's knowledge of subject, organization of content,

preparation) almost always appears near the top of the lists (Crawford & Bradshaw, 1968; Downie, 1952; Gadzella, 1968; Guthrie, 1954; Musella & Rusch, 1968; Smith, 1944). Stimulation of thought appears in a few (Musella & Rusch, 1968; Smith, 1944). Except in Guthrie's studies, the arousal of student interest is not mentioned, but instead, the teacher's interest in the subject is listed (Downie, 1952; Gadzella, 1968).

In comparing the University of Pittsburgh and University of Washington results with the few factor-analytic studies in which factor relationship to overall teaching effectiveness has been examined (Coffman, 1954; Hildebrand et al., 1971; Isaacson et al., 1964; Solomon, 1966), the greatest agreement is with the findings of Coffman and Isaacson et al. Two of the three factors Coffman found to be related to general estimate of teaching effectiveness can be interpreted as clarity and as the arousal of student interest, though he chose other names. (His third factor is not well defined and leaves some question regarding his interpretation of it as "verbal fluency.") In addition, his evidence that personal appearance and punctuality in meeting classes are not related to students' general estimate is borne out in the low rank ordering of these items in the University of Washington study.

Isaacson et al. (1964) indicate that the first of their six factors correlates most highly with "all-around teaching ability." The items describing this factor are as follows: (*a*) Material is put across in an interesting way; (*b*) The intellectual curiosity of the students is stimulated; (*c*) Things are explained clearly; and (*d*) The teacher is skillful in observing student reactions. The last item, (*d*), may be a contributor to clarity, since the teacher who is sensitive to class reactions may perceive that an explanation is not being understood and may, consequently, alter his approach or go over the material again.

There is obvious agreement with the factor described by Hildebrand et al. (1971) as "organization/clarity" and possible overlap with another labeled "dynamism/enthusiasm," though this latter conclusion needs examination. They reported that a rating item constructed to measure this factor correlates more highly with overall teaching effectiveness ratings than do the other four factors (organization/clarity, analytic/synthetic approach, instructor–group interaction, instructor–individual interaction). The actual wording reported for their (Hildebrand et al., 1971) dynamism item is as follows: "Enjoys teaching, is enthusiastic about his subject, makes the course exciting, and has self-confidence [p. 23]."

It is interesting to note that the phrase "makes the course exciting" was *not* among the items factor analyzed (Hildebrand et al., 1971, p. 8, Table 3), or in the factor results (p. 18, Table 5). In this author's opinion, "making the course exciting" does not represent the factor "dynamism" but rather is related to making the course interesting. Certainly it represents something achieved by the teacher, not a personality dimension. It is entirely possible that the insertion of this phrase in an item designed to measure dynamism accounts for its high correlation with overall effectiveness.

Comparison with Solomon's (1966) results is not appropriate for a number of reasons. His rating categories were based on quantity or frequency (never, rarely, sometimes, frequently, always) rather than comparative quality; many of his items are sharply different in nature from those used in most studies (Percentage of time teacher speaks; Amount of teacher factual speech; Number of unsolicited student comments), thus, his factors are not very similar to those found by others; and finally, his sample consists of adult, evening school classes from five different schools treated as one population.

A major difference between the results of other investigators and those of the University of Washington and University of Pittsburgh studies is in the weight given to personality dimensions in the teacher. Asking students to describe good teachers resulted in frequent mentions of qualities such as sympathetic, friendly, enthusiastic, and energetic. Friendliness and sympathetic manner are low in the University of Washington

ranking, while the item "Shows interest and enthusiasm in his subject" ranked ninth. Such personality qualities were omitted in the University of Pittsburgh study. The high multiple correlations in the University of Pittsburgh results and in those of Langen and Sorenson (1963) show that personality characteristics are not necessary in the prediction of student overall ratings of teaching. Enthusiasm on the part of an instructor may contribute to arousing students' interest in the subject, but there are obviously other ways in which the same goal can be achieved.

When asked what they consider the important characteristics of a good teacher, students may tend to generalize from those qualities that they have valued in other personal relationships to interactions with their college teachers. However, in evaluating their instructors, students give high overall ratings to some teachers who are not judged to be dynamic, enthusiastic, sympathetic, or friendly. They may be quiet, scholarly, and a little aloof, or they may have many other combinations of personality traits. (This is not to say that personality factors do not play a part in the methods a teacher chooses or is able to use successfully.)

From the findings of the University of Pittsburgh and University of Washington studies, it can be concluded that the kind of teaching evaluated most highly by students is teaching which they judge to be clear in exposition, which arouses or broadens their interests, and which motivates or stimulates them to intellectual activity. These are broad objectives that can be achieved through the use of different teaching methods, in classes of different sizes, and by teachers with different personalities. Such objectives are far from easy to attain and are perhaps more difficult to achieve in some subject areas than in others. But because they represent changes in the learner, they can become goals to be added to those of the instructor for his or her students. Awareness of what components of teaching students weight most highly should improve our interpretation of their evaluations and provide direction for the construction of rating instruments.

REFERENCES

Bendig, A. W. A factor analysis of student ratings of psychology instructors on the Purdue Scale. *Journal of Educational Psychology*, 1954, **45**, 385–393.

Burket, G. R. A study of reduced rank models for multiple prediction. *Psychometric Monograph*, 1964, No. 12.

Coffman, W. E. Determining students' concepts of effective teaching from their ratings of instructors. *Journal of Educational Psychology*, 1954, **45**, 277–286.

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, 1971, **41**, 511–535.

Crawford, P. L., & Bradshaw, H. L. Perception of characteristics of effective university teachers: A scaling analysis. *Educational and Psychological Measurement*, 1968, **28**, 1079–1085.

Downie, N. W. Student evaluation of faculty. *Journal of Higher Education*, 1952, **23**, 495–496.

French, G. M. *College students' concept of effective teaching determined by an analysis of teacher ratings.* (Doctoral dissertation, University of Washington) Ann Arbor, Mich.: University Microfilms, 1957, No. 21-203.

Gadzella, B. M. College student views and ratings of an ideal professor. *College and University*, 1968, **44**, 89–96.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Guthrie, E. R. *Evaluation of teaching: A progress report.* Seattle: University of Washington, 1954. (Monograph)

Hildebrand, M., Wilson, R. C., & Dienst, E. R. *Evaluating university teaching.* Berkeley, Calif.: Center for Research and Development in Higher Education, 1971.

Hodgson, T. F. *The general and primary factors in student evaluations of teaching ability.* (Doctoral dissertation, University of Washington) Ann Arbor, Mich.: University Microfilms, 1958, No. 59-1224.

Horst, P. Approximating a multiple correlation system by one of lower rank as a basis for deriving more stable prediction weights. In P. Horst (Ed.), *The prediction of personal adjustment.* New York: Social Science Research Council Bulletin, 1941, **48**, 437–444.

Horst, P. A generalized expression for the reliability of measures. *Psychometrika*, 1949, **14**, 21–32.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 1933, **24**, 417–441, 498, 520.

Isaacson, R. L. et al. Dimensions of student evaluations of teaching. *Journal of Educational Psychology*, 1964, **55**, 344–351.

LANGEN, T. D., & SORENSON, R. C. *The influence of ten selected teaching traits on various aspects of course outcomes.* Seattle: University of Washington, 1963. (Mimeo)

LEIMAN, J. M. The calculations of regression weights from common factor loadings. Unpublished doctoral dissertation, University of Washington, 1951.

MCKEACHIE, W. J. *Teaching tips.* (6th ed.) Lexington, Mass.: D. C. Heath, 1969.

MUSELLA, D., & RUSCH, R. Student opinion on college teaching. *Improving College and University Teaching,* 1968, **47,** 18–22.

REHBERG, R. A., ROBERTS, A., & VANDAMENT, W. *Evaluation of university teaching performance via systematic student questionnaire.* Binghamton: State University of New York, 1970. (Mimeo)

REMMERS, H. H., & BAKER, P. C. *Manual of instructions for Purdue Rating Scale for Instructors.* Lafayette, Ind.: Purdue University, Division of Educational Reference, 1952.

RILEY, J. W., JR., RYAN, B. F., & LIFSHITZ, M. *The student looks at his teacher.* New Brunswick, N. J.: Rutgers University Press, 1950.

SMITH, A. A. College teachers evaluated by students. *Sociology and Social Research,* 1944, **28,** 471–478.

SOLOMON, D. Teacher behavior dimensions, course characteristics, and student evaluations of teachers. *American Educational Research Journal,* 1966, **3,** 35–47.

(Received July 17, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 386–389

EFFECTS OF TEACHER SEX AND STUDENT SEX ON THE EVALUATION OF COLLEGE INSTRUCTORS

PATRICIA B. ELMORE[1] AND KAREN A. LAPOINTE

Southern Illinois University at Carbondale

This study assessed the influence of faculty sex and student sex in teacher evaluation. Twenty questions from the Instructional Improvement Questionnaire which directly evaluate instructor performance were analyzed using a two-factor analysis of variance. No interactions between faculty sex and student sex were found. Generally, there were no differences between the mean ratings given male and female faculty by male and female students. However, male instructors did receive higher ratings on "spoke understandably," while female instructors received higher ratings on "promptly returned homework and tests." In addition, female students rated instructors higher on "specified objectives of the course."

For at least a decade, a question of practical importance to university faculty, administrators, and students as well as parents and state legislators has been, How can college teaching be evaluated? An extensive survey conducted by Gustad (1961) revealed that student ratings were most frequently mentioned as the method of teacher evaluation used by almost 600 colleges and universities.

Once a method or procedure was determined to evaluate college teachers, another question was posed: Do particular characteristics of the instructor or the students affect the way the instructor is evaluated by the students in his class? Such characteristics as the sex of the faculty member being evaluated and the sex of the student doing the evaluation may affect the ratings given.

Studies which have attempted to assess the influence of faculty sex on evaluations have found conflicting results. Two investigations (Elliott, 1950; Lovell & Haner, 1955) found no significant differences between male and female faculty.

In studies where differences have been found, investigators are reluctant to place much emphasis upon them. Spencer (1969) reported that at the University of Illinois there is a tendency for women to receive higher ratings. He concluded, however, that there was no meaningful relationship between sex and most Illinois Course Evaluation Questionnaire items. Similarly, Heilman and Armentrout (1936) found higher mean ratings for women than for men, but the differences were not statistically significant. In another study, Downie (1952) found women instructors received higher ratings than men in the extent to which they brought new books and authors into the classroom.

Results from studies examining the effect of student sex are as conflicting as the studies related to faculty sex. Two studies (Goodhartz, 1948; Isaacson et al., 1964) found no differences between faculty ratings made by male and female students. Bendig (1953), however, in a factor analysis of the Purdue Rating Scale for Instruction found two factors: instructor competence and instructor empathy. There were no significant differences due to student sex when both factors were considered together. When instructor competence was considered alone, there was a significant sex difference but the direction of the difference was inconsistent.

A couple of studies have found specific differences between men and women students. Bendig (1952) found that women students rated their instructors (men) signifi-

[1] Requests for reprints should be sent to Patricia B. Elmore, Testing Center, Washington Square, Building C, Southern Illinois University, Carbondale, Illinois 62901.

cantly lower than the male students rated them, whereas Elliott (1950) found that women students tended to give higher ratings in "presentation of the subject matter" than male students.

Recently, studies involving faculty sex, student sex, and characteristics of the instructor, course, or teaching method have found significant sex differences. Carney and McKeachie (1966) found that women rate life-oriented topics in psychology significantly higher and science-oriented topics in psychology significantly lower than men.

In addition, McKeachie, Lin, and Mann (1971) reported that in four out of five studies relating student sex, achievement test results, and factors obtained from student ratings, instructors who received higher ratings in structure and feedback were more effective with women than with men students.

The purpose of the present study was to determine if women faculty received significantly different ratings than men faculty and if this was related to the sex of the student doing the rating.

Method

Students at Southern Illinois University at Carbondale evaluated 1,474 courses during 1971. From this data base, courses were matched on the basis of course number and sex of instructor. A matched sample of 38 pairs of courses evaluated by 1,607 students was obtained. Of these 1,607 students, complete data was available for 1,259 students. The courses included in this study represent varied departments and colleges on the Carbondale campus.

The questionnaire that was administered to the students was the Instructional Improvement Questionnaire.[2] This questionnaire consisted of five parts: Part I, instructor evaluation (Items 1 through 20); Part II, evaluation of course (Items 21 through 40); Part III, strengths and weaknesses (Items 41 through 60); Part IV, research data (Items 61 through 72); and Part V, optional items (Items 73–100). The responses to Part I, which required the students to evaluate different aspects of the instructor's performance on a five-category scale, were used for analysis (see Figure 1).

Results

The students' responses to each of the twenty items were submitted to a two-factor (Instructor Sex × Student Sex) analysis of variance. Due to the extreme negative skewness of the data, the level of significance chosen was .01.

The results of the analysis of variance revealed no statistically significant interaction between faculty sex and student sex on any of the 20 items.

Only two significant differences were found between male and female faculty. Male instructors received significantly higher mean ratings on Item 6, "spoke understandably," than female instructors, while women received significantly higher ratings on Item 15, "promptly returned homework and tests."

Only one significant difference between male and female students emerged. Female students rated instructors higher on Item 13, "specified objectives of the course," than did male students.

Discussion and Conclusions

The present study found no interaction between faculty sex and student sex, and in general, there seemed to be few meaningful differences between male and female faculty. Only two differences, that men spoke understandably and that women more promptly returned homework assignments and tests, were significant. It may be that men on the average project their voices more effectively than women or that the tonal quality of their voices is more pleasing. Women may have more concerns over efficiency than men and, hence, pay more attention to returning tests and homework. Neither of these differences, however, seems to be practically significant.

Similarly, there seems to be little difference between male and female students. The only significant difference found was that women students tended to rate their instructors higher on specifying objectives of the course. Recently, McKeachie et al. (1971) have shown that instructors rated high on structure were more effective with women than with men students. It may be that structure as reflected in organization is more salient and valuable to women than men students.

[2] Copies are available upon request from the senior author.

PART I: INSTRUCTOR EVALUATION: ITEMS 1 THROUGH 20

DIRECTIONS: The following twenty phrases relate to college-level teaching. Evaluate how your instructor did in each of these aspects of teaching by selecting the one response option (A through E below) that comes closest to your judgment.

RESPONSE OPTIONS: OMIT ITEMS THAT DO NOT APPLY

A. Exceptional or outstanding performance
B. Very Good performance
C. Good performance, all that I would normally expect in college-level teaching
D. Weak performance, instructor should be aware of some opportunity for improvement
E. Improvement definitely needed

1. Prepared for class
2. Made clear assignments
3. Set clear standards for grading
4. Graded fairly
5. Knew if students understood him
6. Spoke understandably
7. Answered impromptu questions satisfactorily
8. Showed an interest in the course
9. Gave several examples to explain complex ideas
10. Accepted criticism and suggestions
11. Increased your appreciation for the subject
12. Was dependable in holding class as scheduled
13. Specified objectives of the course
14. Achieved the specified objectives of the course
15. Promptly returned homework and tests
16. Showed an interest in students
17. Made assignments that helped you understand the course
18. Was available outside of class
19. Encouraged participation of students
20. In general, taught the class effectively

FIGURE 1. Part I, instructor evaluation: Items 1 through 20 on the Instructional Improvement Questionnaire.

The present results were consistent with previous findings in which the majority of studies found no significant or practical differences between ratings of male and female faculty or between male and female students' ratings of faculty.

Many variables, such as class size, teaching method (lecture versus discussion), course content, rank of instructor, and student class, are uncontrolled in a large evaluation study like this. This may tend to mask differences which are present. Further experimentation in which class size and instructor rank are controlled is being planned.

Sex differences do seem to occur in studies which assess the interrelationships among student sex, faculty sex, and student or instructor characteristics. Further research is needed to clarify the relationship of student sex, faculty sex, and such characteristics as teacher warmth, structure, and feedback. If sex differences are to be found, it seems that they will be discovered in relationship with these other variables.

REFERENCES

BENDIG, A. W. A preliminary study of the effect of academic level, sex, and course variables on student rating of psychology instructors. *Journal of Psychology,* 1952, **34,** 21–26.

BENDIG, A. W. Student achievement in introductory psychology and student ratings of the competence and empathy of their instructors. *Journal of Psychology,* 1953, **36,** 427–433.

CARNEY, R., & MCKEACHIE, W. Personality, sex, subject matter, and student ratings. *Psychological Record,* 1966, **16,** 137–144.

DOWNIE, N. W. Student evaluation of faculty. *Journal of Higher Education,* 1952, **23,** 495–496.

ELLIOTT, D. H. Characteristics and relationships of various criteria of college and university teaching. *Purdue University Studies in Higher Education,* 1950, **70,** 5–61.

GOODHARTZ, A. S. Student attitudes and opinions relating to teaching at Brooklyn College. *School and Society,* 1948, **68,** 345–349.

GUSTAD, J. W. Policies and practices in faculty evaluation. *Educational Record,* 1961, **42,** 194–211.

HEILMAN, J. D., & ARMENTROUT, W. D. The rating of college teachers on ten traits by their students. *Journal of Educational Psychology,* 1936, **27,** 197–216.

ISAACSON, R. L. et al. Dimensions of student evaluations of teaching. *Journal of Educational Psychology,* 1964, **55,** 344–351.

LOVELL, G. D., & HANER, C. F. Forced choice applied to college faculty rating. *Educational and Psychological Measurement,* 1955, **15,** 291–304.

MCKEACHIE, W. J., LIN, Y., & MANN, W. Student ratings of teacher effectiveness: Validity studies. *American Educational Research Journal,* 1971, **8,** 435–445.

SPENCER, R. *Some dimensions of the Illinois Course Evaluation Questionnaire.* (Res. Rep. 303) Urbana: University of Illinois, Office of Instructional Resources, September 1969.

(Received July 25, 1973)

ANNOUNCEMENT

In addition to the members of the Advisory Board, the *Journal of Educational Psychology* asks other psychologists to aid in the review of manuscripts. In this way we have been able to achieve a broader review base. The Editors wish to thank the following people for their generous assistance to the Journal.

ALEXANDER W. ASTIN, *American Council of Education*

BERNARD BEER, *E. R. Squibb & Sons, Inc., New Brunswick*

DARYL J. BEM, *Carnegie-Mellon University*

MARILYNN B. BREWER, *University of California, Santa Barbara*

EDWARD A. CHITTENDEN, *Educational Testing Service*

FRANK COSTIN, *University of Illinois at Urbana-Champaign*

RUTH B. EKSTROM, *Educational Testing Service*

GERARD C. FANELLI, *Newark State University*

NORMAN T. FEATHER, *University of South Australia*

ROBERT A. FELDMESSER, *Educational Testing Service*

LEONARD S. FELDT, *University of Iowa*

NORMAN FREDERIKSEN, *Educational Testing Service*

NORMAN E. GRONLUND, *University of Illinois at Urbana-Champaign*

HAROLD O. GULLIKSEN, *Educational Testing Service*

PETER GUMPERT, *Teachers College, Columbia University*

RICHARD A. HANSEN, *City University of New York*

JEROME KAGAN, *Harvard University*

PHYLLIS KATZ, *City College of New York*

ROBERT MILWARD, *Brown University*

JOHN NOLAN, *Teachers College, Columbia University*

JOAN PRENTICE, *Indiana University*

JOSEPH SCANDURA, *University of Pennsylvania*

THOMAS SHUELL, *State University of New York at Buffalo*

ARTHUR STAATS, *University of Hawaii*

SIGMUND TOBIAS, *City University of New York*

E. P. TORRANCE, *University of Georgia*

R. D. TENNYSON, *Florida State University*

JAMES TURNURE, *University of Minnesota*

DANIEL WHEELER, *University of Cincinnati*

Journal of Educational Psychology
1974, Vol. 66, No. 3, 390-405

CHANGING TEACHER AND STUDENT BEHAVIOR: AN EMPIRICAL INVESTIGATION[1]

THOMAS L. GOOD[2]
University of Missouri—Columbia

JERE E. BROPHY
University of Texas at Austin

Teacher behavior toward two different groups of target students was altered by presenting teachers with information about their previous interaction with the target children. About 40 hours of pretreatment and 40 hours of posttreatment data were collected in first-grade classrooms drawn from three different schools, using the Brophy-Good dyadic observation system. Treatment procedures markedly altered both quantitative and qualitative aspects of teacher interaction with target students, and in addition, the behavior of the target students changed reciprocally. Changes in specified teacher behavior toward target children were accompanied by many additional changes in unspecified teacher behavior toward both target and nontarget children. The effects of these behavioral changes on nontarget children were generally positive or at least nondetrimental.

Classroom life is an uneven affair. Some students receive much more teacher contact than others (Good, 1970; Horn, 1914; Jackson & Lahaderne, 1967; Jones, 1971; Kranz, Weber, & Fishell, 1970; Mendoza, Good, & Brophy, 1972; Sikes, 1971). Also, some students receive qualitatively superior teacher treatment (Brophy & Good, 1970b; deGroat & Thompson, 1949; Good & Brophy, 1972; Rist, 1970; Rowe, 1969; Silberman, 1969). As the authors have noted previously (Good & Brophy, 1971), when investigators have looked for differential teacher behavior toward students differing in achievement, sex, or socioeconomic status, they have consistently found it. Low-achievement students, for example, usually receive considerably less opportunity to respond than high-achievement students. Thus, different students regularly receive differential treatment from their teachers, and at times such teacher behavior is inappropriate. However, little is known about what feedback procedures help teachers to change their behavior toward selected students and how such changes affect both target and nontarget students.

Seemingly, the easiest way to change teachers' behavior would be to make them more aware of it through feedback. However, teachers often reject feedback because they do not agree with the criteria the supervisor uses (McNeil, 1971). Consequently, teachers usually do not change their teaching along the lines suggested, and they sometimes even deliberately move in the opposite direction (Tuckman & Oliver, 1968).

In response to this difficulty, a model was developed for presenting information to teachers about their behavior in nonthreatening ways (Good & Brophy, 1970). One purpose of this study was to assess the applicability of this model. A single interview was selected as the treatment, because, in addition to testing the model, we wished to design a strategy that could be used to give teachers feedback after observing them. Details on the strategy are provided in the Treatment section. A third purpose of

[1] This research was supported in part by Grant MH 17907-01 from the National Institute of Mental Health, U. S. Public Health Service, Thomas L. Good, principal investigator. The authors gratefully acknowledge the coding assistance of Candy Chazanow, Carolyn Evertson, Suzi Good, Teresa Harris, and Sue Jones and acknowledge the assistance of Susan Florence, Pat Hollowell, and Sherry Kilgore for typing the manuscript.

[2] Requests for reprints should be addressed to Thomas L. Good, Center for Research in Social Behavior, 111 East Stewart Road, University of Missouri, Columbia, Missouri 65201.

the study was to assess effects on *both* target and nontarget students. If the treatment resulted in changed teacher behavior, would this lead to reciprocal changes in student behavior? Behavior modification studies have demonstrated repeatedly that consistent teacher behaviors lead to predictable student responses. However, behavior modifiers and researchers in general have not typically investigated the effects of a treatment on *nontarget* students. Most classroom behavior modification studies have been case studies in which only one or a few target students were observed. If the classroom ecology is to be disturbed, it is important to assess how changes in teacher behavior affect *all* students. For example, if teachers are requested to have more contacts with certain students, does this reduce their contacts with other students? Conversely, perhaps the call for more frequent interaction with certain students has a radiation influence that results in the teacher having more individualized contacts with all students.

Withall (1956) provided information and suggestions to teachers on student participation rates. The teachers dramatically increased the participation level of students who had been low participators. Furthermore, a radiation effect was observed: teachers' interaction rates with nontarget students also rose significantly. These data strongly suggest that treatment effects may radiate to nontarget students, and they indicate the need for more research on unanticipated treatment effects. The call for teachers to behave in *qualitatively* different ways also needs empirical investigation, especially concerning the possibility of disordinal effects. For example, perhaps teachers asked to praise more and to criticize less when dealing with target students will do this with all students. Other instructions may have an opposite effect. Teachers asked to stay with target students when they fail to respond correctly (by probing, repeating the question, providing clues) may begin to deal more expediently with nontarget students (give up easily, call on someone else).

Turner, Foa, and Foa (1971) point out that some resources, like money, are potentially exhaustible. If one continues to give money, it eventually runs out. Other resources are like love: one can continue to give without extinguishing the source. More attention is needed to pinpoint teacher behaviors that can be increased toward target children without reducing the expression of those same behaviors to other students. To summarize, the two major purposes of this study were as follows: (*a*) to see if a simple feedback strategy derived from a more general model would change teacher behavior and (*b*) to observe the effects of any changes in teacher behavior on *both* target and nontarget students.

Treatment Rationale

It was hypothesized that feedback on observational data would be more acceptable to teachers than the feedback they usually get from supervisors. As noted elsewhere (Good & Brophy, 1970), supervisor comments too often boil down to "your way is wrong, do it my way," leading teachers to become defensive. The treatment attempted to minimize this problem in two ways. First, it was based on objective observation of a large number of discrete interactions, rather than upon brief impressions. In fact, the observers in each teacher's classroom were not involved in the treatment, so that advice would be based strictly on objective data. Interviews were conducted by the authors, working from data sheets turned in by the observers. This stress on objective data was designed to reinforce the acceptability and credibility of the suggestions given to teachers.

A second advantage resulted from the use of the individual student as the focus of analysis. By tabulating the teacher's interaction with each different student separately, it was possible to show that teachers were teaching some students appropriately and others inappropriately. In making suggestions for improvement, the authors were in effect saying "you are doing a fine job with Mary, now try to do the same kinds of things with Jane." This is much less threatening and more acceptable than "your way is wrong, do it my way."

Method

The treatment was confined to a single interview to see if teachers could change in response to feedback alone, without any retraining or continuing supervision. Earlier work had convinced the investigators that much inappropriate treatment of students occurs partly because teachers are not aware of what they are doing. Few teachers deliberately and consciously give up on a student or treat him with rejection and discouragement rather than with patience and encouragement. Instead, they gradually drift into a pattern in which they are reacting to certain students inappropriately without realizing it. If this is true, it follows that a treatment which simply provides teachers with *information* to make them aware of what they are doing should be enough to make them able to change.

Treatment Groups

The treatment was focused on two types of students in each classroom. The first (low-participant group) were students with low rates of interaction with teachers. These students seldom volunteered to answer questions or initiated interactions. Correspondingly, the teachers did not seek them out very frequently. Thus, the low-participant group had low rates of interaction with teachers, both because they avoided the teachers and because the teachers avoided them.

A second group (the extension group) was identified on the basis of their teacher's willingness to provide them with a second response opportunity when they failed to succeed on the first. Teachers usually did not persist in seeking responses from extension students if they failed on their first opportunity. Instead, the teachers would give them the answer or call on someone else. These students were labeled the "extension" group because the advice given to teachers was to *extend* their interactions with them in failure situations.

Each classroom showed extreme variability on these measures, so that low-participant and extension students could be identified in each class. There were also general differences from one teacher to the next. For example, one teacher stayed with her students over 50% of the time following an initial failure, while another stayed less than 20% of the time. However, within each class, there were some students with very high percentages and others with very low ones. The situation with low participants was similar. There were large differences between classrooms but even larger differences within classrooms, so that a group with strikingly low interaction frequencies could be identified in each class.

In each classroom at least two students, and more typically three or four, were identified for each treatment group. The number was dependent upon the data. If only two students stood out as having strikingly low rates of interaction with the teacher, these students alone comprised the low-participant group. In another class where four students had strikingly low rates, the low-participant group would contain all four. This flexibility allowed us to be sure that each student included in a group was appropriately classified and that our suggested treatment was, therefore, appropriate for him.

A total of 21 low participants and 28 extension students were identified in the eight classrooms. The low-participant group included 14 girls and 7 boys, while the extension group included 8 girls and 20 boys. The predominance of girls in the low-participant group was expected, since girls typically have fewer interactions with teachers than boys and are generally less salient in the classroom (Brophy & Good, 1974).

Perhaps the extension group contained more boys than girls because of the sex difference favoring girls in achievement at this grade level. To the extent that teachers' rates of staying with students are affected by the probability that the second question will be answered successfully, this sex difference in achievement may have been a factor. Both groups included students from all achievement levels.

Treatment Interview

The data were analyzed between semesters, and interviews were scheduled early in the second semester. Each teacher was seen individually. For each interview, a list had been prepared containing the names of *four* groups of students. On the left side were the names of the low participants and, under them, the names of the extension students. On the right side, across from each respective treatment group, were the names of two *contrast groups*.

The contrast groups were students that the teachers were treating appropriately in situations where they were treating the treatment groups inappropriately. The contrast children were included for two reasons. The first was to stimulate each teacher's thinking and help her to "see" her own behavior more clearly. Hopefully, the opportunity to compare each treatment group with its contrast group would help the teacher gain insight into the reasons for her differential treatment of these two groups. The other reason for including contrast groups was our interest in providing acceptable and nonthreatening feedback. The contrast groups made it possible to point to students with whom the teacher had been doing an especially good job. Further, in suggesting changes, it was possible to say "you are already doing it with these contrast students; now try to do it with these other students, too."

Teachers were first shown the list containing the names of the four groups and asked to respond to it. They were puzzled by this request, and most made vague responses. A couple indicated that the low participants included "some of the shyest students in the class." Otherwise, the teachers were unable to indicate common characteristics within groups or common differences between contrast groups.

The teachers were then told how the groups had been selected and how they had been behaving differentially toward them. About half of the teachers were aware of their low rates of interaction with *some* low participants. Their responses included "Deborah doesn't need help and doesn't want to be bothered by anyone. She is very independent"; "Carla is very shy. She is extremely embarrassed when I call on her and she can't answer. I don't want to embarrass her." When a teacher was aware of her low rate of interaction with a student, the student was typically a shy one that the teacher was afraid of embarrassing, although a few "independent" students were included here also. The low-participant group also contained "unknowns" in addition to shy and independent students. Teachers usually were not aware of their rates of interaction with such students, nor particularly aware of these students generally. They responded to questions about them with statements like "you know Sam there ... I guess I don't interact much with him. I don't know him very well."

Attempts to get the teachers to compare the low-participant group with its contrast group yielded no clear differences, although a few trends were evident. Contrast students were more likely to be described as independent than as shy, and they were not seen as likely to become embarrassed or upset if they could not answer a question. Thus, they contrasted well with the shy students in the low-participant group but not with the "independents" or the "unknowns."

After discussion, the teachers were asked to increase their interactions with low participants by calling on them more frequently and initiating more private interactions with them. The teachers readily agreed to this with regard to the "unknowns," but many had reservations when it came to the independent students (they don't need it—leave well enough alone) and especially to the shy students (they will become embarrassed and upset).

One response here was that the teachers' concerns were understandable but that these students' willingness to respond in public situations could be improved without endangering their security. It was suggested that inhibitions can be eliminated only with practice and exposure to public response opportunities, especially through low-key opportunities that result in successful student response. Teachers were encouraged not only to seek out the low participants more often for private contacts but also to deliberately call on them to increase their public response opportunities. We also reminded the teachers that they did not have to force a student to respond if he did start to become flustered; they could help him by giving the answer or providing a clue. After varying amounts of discussion, the teachers agreed to try to increase private and public contacts with low participants, although a few teachers had reservations about a few students. The latter were mostly high-achieving, independent types, and the teachers felt (with some justification) that their contact patterns were already optimal.

Similar treatment procedures were followed for the extension group. The teachers could identify no common characteristic of the extension students and no common differences between them and their contrast group. Also, *none* of the teachers had any awareness of their tendency to give up easily with the extension group or to be especially persistent with the contrast group. Each teacher was surprised and mystified by the data on this variable. A few said things like "that child makes me nervous," about individual extension students, but they were unable to name any attribute that was common to all or most of them. *Apparently, unknown individual differences in the students condition the teacher into quite predictable behavior patterns, but this is done without any awareness on her part.*

All the teachers readily agreed to become more persistent in seeking responses from the extension students, either by repeating the question or by providing help in the form of a clue or a new question. Teachers expressed more interest in changing this behavior (perhaps because they were so totally unaware of this aspect of their behavior) than they had in changing their behavior toward low participants.

Data Collection

Data were collected in eight out of nine first-grade classrooms that were already involved in a larger study of the relationships between teachers' performance expectations and their behavior toward different children (Brophy & Good, 1974). Originally, there were three classes studied in each of three types of schools: upper-middle-class white, lower-class white, and lower-class black. However, one teacher at the upper-middle-class, white school moved out of town after the first semester. Teachers were told that the investigators were interested in observing differences in the classroom behavior of children who varied in achievement. In late September, each teacher supplied a list ranking her children in order according to the levels of achievement she expected from them. Teachers were also asked to rank their students at two other times in the year.

Sixteen 2½ hour observations were made in each classroom with the Brophy-Good dyadic interaction system (Brophy & Good, 1970a). The resulting data pool provided information on the teacher–child interaction patterns of 259 children, based on 40 hours of classroom observation taken on 16 different days during a 3-month period. After baseline data were collected, the treatment was administered and then an additional 3 months of behavioral data were collected.

During September, pairs of observers worked in each classroom to establish reliability (80% agreement) and to desensitize the teachers and children to their presence. After reliability was established (procedures are detailed in Brophy & Good, 1970a), the observers began to work singly. Ob-

servers did not see the teachers' achievement rankings and did not know which students had been singled out for special study.

Data Analysis

The raw data were first converted into measures designed to control for absences and to allow direct comparison among children in the same room. Frequency counts were converted to means, dividing each child's totals by the number of observations for which he was present. Other measures were percentage scores compiled according to the procedures detailed in Brophy and Good (1970a).

The data for each class were then standardized ($M = 0$, $SD = 1$) to set the eight classes on a common scale and to eliminate variance due to teacher or class differences. Analyses of variance were then obtained from these standardized score distributions to test for significant changes in teacher interaction with target students. First, two series of one-way analyses of variance over repeated measures (pre- and posttreatment) were performed, in which the means on the standardized teacher–student interaction measures for each target group, respectively, were compared with the means for all other children. Thus, these analyses compared the scores of 21 or 28 children with those of the remaining 238 or 231, respectively. In addition, repeated measures analyses of variance were performed separately for each classroom, using nonstandardized means and percentage scores. These data reveal the extent to which treatment effects generalized across teachers and radiated to include nontarget children.

Results

Achievement Expectations

To see whether the treatment affected the teachers' achievement expectations for target students, teachers' rankings collected in mid-November were compared with the set collected in March, after the treatment interview. These comparisons showed very little change in the rankings. Thus, despite the considerable changes in teacher–child interaction patterns to be described in the next section, the treatment did not affect the teachers' achievement expectations for the target students.

Teacher–Child Interaction Effects Involving Target Groups

Comparison of teacher–child interactions involving target students before and after treatment showed clear-cut treatment effects for both target groups. The data are summarized in Table 1.

Extension Group

The most notable changes in teacher behavior with the extension group occurred on the measure of staying with students following an initial failure. Despite a marked tendency to give up easily with these students in the first semester, in the second semester the teachers stayed with these students just as often as they did with their classmates. Thus, the treatment brought the extension students from a state of clear disadvantage to a state of parity with their classmates.

There were also other changes in the teachers' treatment of extension students. The teachers called on them more frequently and initiated more contacts with them during the second semester, and they praised them more frequently than they had in the first semester. In addition to the increased praise, an improvement in the ratio of behavioral warnings to behavioral criticisms also suggests an improvement in the teachers' attitudes toward extension students. In the first semester, the teachers were prone to respond to misbehavior by extension students with criticism. In the second semester, they more often merely warned the students. However, even after treatment, extension students were more likely than their classmates to be criticized for failures to respond or for giving wrong answers.

The change in teacher behavior toward extension students produced no notable changes in the behavior of the extension students toward the teachers. The only significant change was on the measure of wrong answers over wrong answers plus no response. The extension students had been higher on this measure prior to treatment, but they were equal to their classmates after treatment. This may reflect a decrease in the extension students' willingness to take a guess. However, given the great increase in the teachers' frequencies of staying with these students, the change probably occurred primarily because the teachers began asking them more questions that they could not answer. On measures of percentage of correct answers and frequencies of reading errors, the extension students

were similar to their classmates during both semesters. The same is true for their frequency of behavior contacts with the teacher; this was the case even though most of these students were boys and despite the higher than average teacher criticism of this group. Apparently there was something about these students that made their teachers hypercritical of them, but it was *not* a higher frequency of misbehavior on their part.

In summary, the teachers were able to significantly increase their rates of staying with extension students in failure situations. Following treatment, the teachers stayed with these students just as often as they did with their classmates. They also sought them out for contacts, praised them more often, and were less prone to criticize them. However, extension students still received more criticism when they failed to respond correctly (but not more behavior criticism). There were no notable effects of the treatment upon the behavior of the extension students themselves.

Low-Participation Group

The treatment was also successful in changing the teachers' treatment of the low participants. Seven of eight measures of quantity of teacher–student contacts increased with the low-participant group; most of these, significantly. On some measures, the low participants achieved parity with their classmates, while on others they still remained below average. In other words, the teachers significantly increased their interactions with low participants following the treatment, but this group still had fewer contacts with the teachers than did their classmates.

This activity of the teachers produced some complementary behavior on the part of the low-participant students. There was some tendency for them to seek response opportunities and contacts with teachers more frequently after the treatment. Despite this improvement, however, they remained generally below their classmates on these measures.

There were dramatic changes in the rates of teacher praise and criticism of low participants. During the first semester, they had been praised more frequently than their classmates for good behavior but not for successfully responding to questions or completing seatwork. The posttreatment data showed dramatic rises in teacher praise for correct answers and good seatwork. There was also a corresponding drop on most measures of teacher criticism. These differences are probably related to the teachers' concern about pushing these students too hard and about making sure to encourage them when they showed signs of progress. It should be noted, however, that it took the treatment to bring out this concern: during the first semester, the low participants earned special praise only for their behavior (conforming to classroom rules) but not for their success in answering questions or completing seatwork. *Thus, the treatment focused teacher attention more closely on the academic performance of these students.*

The data for the difficulty level of teacher questions directed to the low participants showed mixed results. In general, the teachers asked them more difficult questions following the treatment. Thus, even though most of the teachers expressed fear of embarrassing them, this concern did not cause them to ask the low participants only easy questions during the second semester. The measures of teacher feedback also showed significant and beneficial changes. During the first semester, teachers frequently failed to give feedback to the low participants after responses, both during reading groups and in general class discussions. Also, teachers less frequently gave these students process feedback when they checked their seatwork. These differences disappeared following the treatment, again suggesting that the teachers began to pay more careful attention to the academic responses of the low participants.

To summarize the teacher effects, the treatment produced more widespread changes in the behavior of teachers toward the low participants than in their behavior toward the extension students. The teachers not only called on these students more often and initiated more individual contacts with

TABLE 1

STANDARDIZED SCORE MEANS AND PROBABILITY DATA FROM GROUPS (STUDENTS IN TREATMENT GROUP VERSUS STUDENTS NOT IN TREATMENT GROUP) BY TRIALS (PRETREATMENT VERSUS POSTTREATMENT) ANALYSES OF VARIANCE IN STANDARDIZED CLASSROOM INTERACTION MEASURES

Dependent variable[a]	Extension							Low participation						
	Group *M*				Probability level[b]			Group *M*				Probability level[b]		
	Pretreatment		Posttreatment					Pretreatment		Posttreatment				
	Other	Extension	Other	Extension	Groups (G)	Trials (T)	G × T	Other	Low participation	Other	Low participaton	Groups (G)	Trials (T)	G × T
Student performance indicators														
Correct answers/total answers	.01	.22	−.03	.07	.68	.52	.66	.05	−.04	.02	−.33	.22	.52	.65
Correct answers/total answers	.07	−.20	.01	.09	.54	.86	.13	.01	.19	.07	−.41	.62	.86	.01
Errors per reading turn	−.04	.16	−.07	.14	.21	.71	.99	.00	−.05	−.05	.08	.80	.71	.54
Wrong answers/wrong + no response	−.04	.38	−.02	−.05	.27	.60	.09	.01	.20	−.04	.15	.64	.60	.99
Wrong answer/wrong + no response	.05	.08	.01	.05	.83	.69	.99	.06	−.03	.03	−.12	.54	.69	.82
Frequencies of each type of teacher–student interaction														
Response opportunities (direct, open, call outs)	.02	−.01	.01	.21	.58	.83	.64	.09	−.67	.03	.05	.03	.83	.01
Called on by teacher	−.01	−.07	−.03	.35	.29	.61	.07	.06	−.69	.01	.13	.06	.60	.00
Called out answer	−.03	.28	−.01	.30	.05	.81	.99	.09	−.57	.08	−.33	.00	.81	.59
Reading turns	.06	.01	−.02	.10	.80	.54	.51	.09	−.20	.02	−.14	.18	.54	.66
Work recitation	.05	.15	.01	.34	.15	.95	.30	.10	−.23	.08	−.13	.11	.95	.66
Self-reference questions	−.02	.15	.01	.19	.21	.77	.97	.05	−.36	.06	−.17	.04	.77	.50
Self-reference questions	−.01	.26	.02	.20	.15	.83	.69	.07	−.31	.07	−.17	.08	.83	.62
Self-reference recitations	−.04	−.01	−.03	.41	.04	.63	.07	−.05	.06	.07	−.24	.55	.63	.11
Total teacher-initiated procedure contacts	−.04	.34	−.08	.25	.02	.53	.80	.08	−.52	−.01	−.23	.02	.53	.10
Total student-initiated procedure contacts	−.03	.25	.04	−.09	.68	.86	.03	.06	−.46	.05	−.27	.03	.86	.66
Total teacher-initiated work contacts	−.03	.06	−.04	.13	.60	.98	.70	.06	−.72	.01	−.30	.00	.98	.05
Total student-initiated work contacts	.02	.09	−.01	.01	.80	.50	.78	.11	−.63	.05	−.48	.00	.50	.27
Total behavioral contacts/total work contacts	−.03	.01	.02	−.17	.64	.86	.31	−.02	−.09	.03	−.34	.24	.86	.23

Teacher versus student initiation of contacts														
Direct questions/direct + open questions	.00	−.14	.02	−.05	.51	.80	.76	.01	−.22	.03	−.23	.16	.80	.90
Direct questions/response opportunities	.03	−.24	.01	.03	.55	.81	.27	−.02	.04	.03	−.17	.68	.81	.61
Open questions/response opportunities	.01	.05	.00	.12	.60	.99	.75	−.03	.38	.01	.02	.22	.99	.20
Call outs/response opportunities	−.03	.17	−.04	.33	.08	.91	.52	.04	−.31	.05	−.30	.06	.91	.99
Call outs/total response opportunities	−.05	.41	.00	.25	.02	.86	.59	.08	−.55	.08	−.33	.00	.86	.57
Total student-initiated contact (work and procedure)	.00	.18	.00	−.04	.70	.62	.18	.09	−.60	.06	−.54	.00	.62	.60
Total teacher-initiated contacts (work and procedure)	−.04	.15	−.07	.23	.12	.86	.59	.08	−.78	.00	−.25	.00	.85	.01
Student-initiated work contacts/total student-initiated contacts	.01	−.12	−.06	.23	.57	.91	.13	.01	−.17	.00	−.20	.25	.91	.96
Teacher-initiated work contacts/total teacher-initiated contacts	.02	−.14	.02	.10	.80	.67	.65	.00	−.08	.04	−.06	.61	.67	.93
Student-initiated work contacts/total work contacts	.06	.06	.01	−.02	.92	.50	.90	.09	−.28	.02	−.13	.18	.50	.57
Student-initiated procedure contacts/total procedure contacts	−.02	.09	.05	−.19	.66	.81	.14	.02	−.26	.03	−.09	.26	.81	.58
Type of teacher questions														
Process questions/total questions	.02	.20	.02	−.04	.66	.71	.63	.07	−.18	−.01	.20	.91	.71	.12
Process questions/total questions	.00	.15	−.02	.17	.23	.87	.87	.06	−.26	.03	−.15	.13	.87	.62
Choice questions/total questions	−.06	.02	.00	.09	.56	.51	.97	−.08	.24	.01	.05	.27	.51	.64
Choice questions/total questions	−.02	.09	−.07	−.17	.95	.61	.62	.02	−.25	−.09	−.03	.52	.61	.24
Self-reference questions/total response opportunities	−.01	−.01	.01	−.05	.81	.95	.82	.00	−.04	.01	−.11	.62	.95	.78
Self-reference questions/total response opportunities	−.01	.23	.01	.01	.56	.79	.65	.01	.18	.01	−.01	.66	.79	.52
Teacher praise and criticism														
Praise after correct answers/total correct answers	.13	−.20	.04	−.10	.10	.57	.53	.04	.39	−.02	.41	.02	.57	.80
Praise after correct answers/total correct answers	.03	.08	.00	.08	.63	.77	.91	.04	−.01	−.04	.42	.21	.77	.13
Praise after reading turn/total reading turns	−.04	−.14	.02	.02	.69	.56	.70	−.04	−.15	−.04	.55	.11	.57	.02
Praise in student-initiated work contacts/total student-initiated work contacts	.00	−.06	−.04	.20	.51	.95	.31	−.02	.16	.02	−.32	.69	.95	.22
Praise in teacher-initiated work contacts/total teacher-initiated work contacts	−.04	−.04	−.07	.26	.31	.80	.20	−.03	−.19	−.07	.60	.21	.79	.01
Total praise of academic work	−.02	.24	−.04	.26	.05	.89	.87	.05	−.36	−.06	.52	.61	.89	.00
Total behavioral praise	−.01	−.09	.01	−.01	.66	.69	.76	−.06	.30	.02	−.05	.26	.68	.09
Criticism after wrong answers/total wrong answers	.03	−.02	.05	−.23	.30	.83	.51	.05	−.18	.03	−.28	.19	.83	.83

TABLE 1—*Continued*

Dependent variable[a]	Extension							Low participation						
	Group *M*				Probability level[b]			Group *M*				Probability level[b]		
	Pretreatment		Posttreatment					Pretreatment		Posttreatment				
	Other	Extension	Other	Extension	Groups (G)	Trials (T)	G × T	Other	Low participation	Other	Low participation	Groups (G)	Trials (T)	G × T
Teacher praise and criticism—*Continued*														
Criticism after wrong answers/total wrong answers	.12	−.13	.04	.00	.67	.72	.53	.07	.16	.07	−.29	.50	.72	.28
Criticism following no response/total no response	.03	−.08	−.09	.00	.91	.62	.53	.02	−.10	−.06	−.21	.56	.61	.95
Criticism after no response/total no response	−.03	.11	−.13	.34	.03	.70	.28	.01	−.25	−.03	−.27	.25	.70	.96
Criticism at end of reading turn/total reading turns	.02	−.08	.01	.04	.80	.98	.54	.04	−.20	.03	−.15	.16	.98	.83
Criticism in student-initiated work contacts/total student-initiated work contacts	.05	.09	.08	−.10	.66	.97	.52	.09	−.47	.06	−.08	.11	.97	.64
Criticism in teacher-initiated work contacts/total teacher-initiated work contacts	−.04	.12	−.01	.19	.25	.72	.88	.04	−.60	.08	−.55	.00	.72	.98
Criticism after reading errors/total reading errors	−.08	.27	−.06	.30	.01	.85	.95	−.01	−.19	.02	−.23	.19	.85	.84
Total criticism of academic work	−.03	.25	−.04	.31	.04	.99	.77	.06	−.52	.06	−.45	.00	.99	.79
Total behavioral criticism	−.05	.39	−.01	.13	.08	.99	.09	.03	−.21	.02	−.11	.66	.99	.60
Warning/warning + criticism	.08	−.37	−.04	−.07	.10	.60	.15	.05	−.41	−.05	.00	.26	.60	.17
Teacher persistence in eliciting responses														
New questions following right answers/total right answers	−.01	−.24	.03	.17	.76	.67	.17	−.06	.11	.04	.10	.50	.67	.73
New question after right answer/total right answers	.07	−.10	.06	−.27	.06	.72	.55	.08	−.32	−.02	.30	.82	.72	.02
% stays with student following failure	.14	−.72	.03	−.05	.01	.91	.01	−.01	.17	−.01	.26	.27	.91	.81
% stays with student following failure	.20	−.80	.00	.15	.00	.88	.00	.01	.27	.01	.17	.24	.89	.78
% stays with student following reading errors	.05	−.16	−.06	.08	.79	.67	.22	.01	.00	−.03	−.09	.85	.68	.89
Repeats question/repeats + gives help (failure situations)	−.15	−.04	.19	.22	.79	.03	.84	−.15	.04	.19	.37	.59	.03	.99
Repeats question/ repeats + gives help (failure situation)	−.03	.28	.10	−.23	.95	.71	.07	.00	.19	.04	.33	.63	.71	.83
Repeats question/repeats + gives help (after reading errors)	.02	.22	−.15	.24	.24	.54	.70	.04	.19	−.02	−.84	.32	.55	.15

	Level of feedback given to the student													
Process feedback/response opportunities	.06	−.12	.00	−.08	.60	.67	.74	.07	−.26	−.03	.22	.81	.66	.07
Process feedback/product feedback	.18	−.06	−.13	.09	.94	.10	.21	.11	.59	−.08	−.35	.74	.10	.21
Process feedback/total response opportunities	.02	.07	.04	−.11	.68	.93	.51	.04	−.08	.01	.10	.92	.93	.54
Process feedback/product feedback	.07	−.04	−.03	.18	.73	.73	.31	.08	−.28	−.02	.25	.81	.73	.12
Process feedback after reading errors/product feedback after reading errors	.07	−.05	−.01	−.13	.56	.51	.98	.04	.12	−.01	−.21	.74	.51	.56
Process feedback at end of reading turn/total reading turns	.03	−.16	.01	−.04	.26	.97	.55	−.02	.18	.01	−.08	.70	.97	.28
Process feedback in student-initiated work contacts/total student-initiated work contacts	.00	.14	.05	−.29	.51	.76	.11	.03	.01	.01	−.22	.59	.76	.66
Process feedback in teacher-initiated work contacts/total teacher-initiated work contacts	−.05	−.05	.02	.07	.88	.55	.84	−.08	.34	.02	.06	.24	.56	.25
Total process feedback	.01	.04	.02	−.05	.90	.97	.68	.06	−.36	.01	.00	.24	.97	.11
% responses in which teacher gives no feedback	−.04	.40	.00	−.10	.20	.70	.05	−.02	.32	−.05	.29	.03	.71	.98
% no feedback/total response	.02	.16	−.05	.43	.03	.79	.22	−.01	.49	.02	.00	.14	.79	.11
No feedback after reading turn/total reading turns	.00	.11	−.06	.25	.11	.74	.50	.01	.06	−.01	−.09	.90	.74	.68

Note. When .00 appears in the table it does not mean zero. The symbol means that the number is less than .01 (.003, .0018, etc.).

[a] Italicized measures are from interaction during reading groups. Nonitalicized measures are from general class activities.

[b] Probability values reflect the percentage probability that the effect is due to chance rather than to a real group difference. Values of .05 and below are considered statistically significant. The values in the "Groups" column refer to general differences between treatment groups and their classmates, across the two semesters combined. The values in the "Trials" column refer to differences between the first and the second semester, across all students combined. The values in the "G × T" column refer to the Groups × Trials (Semesters) interaction, and therefore reflect the significance of changes in group means from the first to the second semester (i.e., these values reflect the effects of the treatment on the target students).

them, they also asked them somewhat more difficult questions, praised them more frequently, criticized them less frequently, gave them more process feedback, and less often failed to give them feedback. Thus, the treatment seemed to make the teachers generally more aware of and concerned about the achievement of the low participants.

There were also some effects on the students. Most notably, the low participants showed a drop in their percentage of correct answers. It is difficult to say whether this change was good, bad, or indifferent. The changes were general but rather small, especially in reading groups where most of the response opportunities occurred. During reading groups, low participants answered 75% correctly the first semester and 64% the second. Meanwhile, their classmates dropped from 75% to 70%. The drop in percentage correct during general class discussions, however, was much greater: low participants dropped from 71% to 55%, while their classmates only dropped from 70% to 67%. Thus, one result of the teachers' asking low participants more frequent and somewhat more difficult questions was that their percentage of correct responses dropped. Except for errorless learning advocates, most observers would see changes of this magnitude as positive, indicating greater teacher efforts to get the most out of these students (assuming, of course, a generally supportive atmosphere).

The posttreatment means for two classrooms dipped to 38% and 31%, respectively, on the measure of percentage of correct answers during general class discussions. In these two instances, the treatment apparently caused the teachers to push too hard. However, the means for the other six classes on this measure and the means for all eight classes on the measure for correct answers in the reading group were in the 50% to 80% range. Thus, although one effect of the treatment was to lower the percentage of correct responses made by low participants, their posttreatment percentages were still comparable to those of their classmates and, judging by the increases in rates of initiation of interactions with the teachers shown by these students, this did not cause fear or embarrassment.

Other changes in student measures were more clearly positive. Although the changes were not nearly so dramatic as those in the data for the teachers, there are suggestions that the low participants became better behaved, more likely to seek response opportunities, and more likely to come to the teacher for help with their work during the second semester. They were still less likely than their classmates to seek out the teacher during the second semester, but they were more likely to do so than they had been before the treatment.

Radiation Effects

The treatment had clearly changed the behavior of the teachers toward both groups. We now turn to the question of how these changes affected the more general classroom ecology. Did the effects radiate? If so, was it to the benefit or the detriment of the classmates of the treatment students? To answer these questions, the raw (unstandardized) means for treatment and nontreatment students for each semester were examined for each separate classroom. These data are presented in Tables 2 and 3.

One factor complicating our analyses was a general change (trials effect) from the first to the second semester on most scores. These changes resulted from changes in the nature of the activities going on in the classrooms at the time and were unrelated to the treatment study itself. For example, there were more reading turns during the second semester than the first. Changes of this sort tended to be constant across all students within a classroom, however, and to be unrelated to treatment changes.

The data regarding radiation of effects for *extension* students generally supported the expectation that there would be some radiation and that the radiation would be beneficial to the whole class. Analyses revealed that seven of the eight teachers showed large gains in their percentages of staying with extension students in failure situations, while one teacher did not change. Of the seven teachers who learned to stay with extension students longer in

TABLE 2

REPEATED MEASURES OVER TWO SEMESTERS (PRE- AND POSTTREATMENT) ON SELECTED INTERACTION VARIABLES BETWEEN EACH TEACHER AND EXTENSION STUDENTS AND NONTARGET STUDENTS

Variable	Classroom															
	1		2		3		4		5		7		8		9	
	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Response opportunity (direct, open, call out)																
Extension	1.24	1.44	1.26	1.13	.36	.33	.71	1.26	1.32	.13	2.08	1.54	1.22	2.63	.75	1.50
Other students	1.80	1.77	1.01	.88	.52	.42	1.12	.88	1.47	.21	.98	.96	1.16	1.85	.93	1.50
Teacher-initiated work contacts																
Extension	.92	.77	1.26	1.66	.73	.36	.47	.08	.65	.13	.82	.35	1.56	1.33	.72	.40
Other students	1.34	.92	1.01	1.08	.81	.40	.68	.20	.61	.06	.77	.33	1.26	1.17	.35	.25
Number of reading turns																
Extension	.68	1.21	.59	.90	.83	1.03	1.00	.54	.93	.63	.85	1.56	1.08	1.49	1.06	1.29
Other students	.72	1.30	.59	.69	.83	1.01	.68	.77	1.11	.54	.76	1.28	1.11	1.26	1.52	1.40
Called on by teacher in reading group																
Extension	2.18	2.36	2.65	1.32	2.74	2.29	1.65	3.43	2.82	3.23	2.17	4.83	2.34	4.28	1.50	3.21
Other students	2.24	1.99	2.61	1.48	3.09	2.16	2.35	3.11	2.85	3.03	1.93	3.99	2.14	3.89	1.59	.78
% stays with student following failure																
Extension	.04	.45	.09	.15							.22	.15	.08	.72	.28	.50
Other students	.35	.51	.28	.39							.43	.18	.23	.60	.30	.29
% stays with student following failure																
Extension	.05	.69	.16	.47	.14	.40	.07	.09	.18	.73	.11	.31	.07	.41	.06	.41
Other students	.31	.64	.28	.32	.34	.24	.26	.27	.48	.42	.30	.39	.25	.58	.28	.35
Student-initiated work contacts																
Extension	.85	.85	.09	.07	.16	.29	.20	.21	.42	.29	1.45	.71	.73	.85	.63	.13
Other students	1.18	1.39	.12	.07	.19	.30	.22	.19	.93	.26	.63	.55	.40	.39	.95	.57
Total praise of academic work																
Extension	9.00	15.60	2.75	1.15	2.20	2.56	6.00	5.22	12.33	11.84	4.00	2.70	17.25	16.00	6.50	3.36
Other students	11.50	15.42	3.00	2.29	1.23	1.20	5.83	8.00	10.24	5.28	2.68	1.58	12.41	13.28	7.43	2.48

Note. Abbreviations: S1 = first semester and S2 = second semester.

[a] Italicized measures are from interaction during reading groups. Nonitalicized measures are from general class activities.

TABLE 3

REPEATED MEASURES OVER TWO SEMESTERS (PRE- AND POSTTREATMENT) ON SELECTED INTERACTION VARIABLES BETWEEN EACH TEACHER AND LOW-PARTICIPATION STUDENTS AND NONTARGET STUDENTS

Variable[a]	Classroom															
	1		2		3		4		5		7		8		9	
	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Response opportunity (direct, open, call out)																
Low participation	1.14	2.79	.60	.90	.27	.24	.59	.87	1.07	.25	.59	.68	.83	2.08	.62	1.49
Other students	1.76	1.62	1.09	.92	.51	.41	1.15	.93	1.50	.20	1.22	1.10	1.21	1.94	.94	1.50
Teacher-initiated work contacts																
Low participation	.99	1.07	.63	1.35	.70	.11	.33	.04	.28	.00	.75	.25	.96	1.43	.15	.19
Other students	1.30	.88	1.08	1.15	.81	.42	.71	.21	.66	.07	.78	.34	1.34	1.16	.43	.28
Number of reading turns																
Low participation	.62	1.11	.56	.54	.42	.46	.70	.99	1.07	.29	.63	1.34	1.05	1.26	1.31	1.63
Other students	.72	1.30	.60	.74	.86	1.06	.72	.70	1.09	.58	.79	1.32	1.12	1.30	1.49	1.36
Called on by teacher in reading group																
Low participation	1.77	2.78	1.79	2.09	1.81	1.29	1.36	3.50	1.90	3.05	1.34	4.58	1.32	3.60	1.55	.73
Other students	2.27	1.99	2.69	1.40	3.12	2.25	2.44	3.08	2.98	3.05	2.05	4.06	2.27	3.99	1.59	1.10
% stays with student following failure																
Low participation	.40	.92	.13	.71							.67	.20	.00	.36	.33	.30
Other students	.29	.47	.26	.31							.35	.17	.22	.64	.30	.32
% stays with student following failure																
Low participation	.22	.96	.20	.39	.52	.32	.50	.37	.69	.17	.10	.35	.19	.70	.17	.25
Other students	.27	.62	.26	.34	.29	.27	.21	.23	.39	.52	.28	.38	.23	.53	.25	.38
Student-initiated work contacts																
Low participation	1.26	1.57	.00	.07	.00	.15	.03	.08	.40	.04	.27	.10	.31	.41	.28	.07
Other students	1.12	1.28	.12	.07	.20	.31	.25	.21	.94	.29	.82	.64	.47	.46	.99	.57
Total praise of academic work																
Low participation	9.50	22.24	1.67	1.70	.50	2.00	4.75	5.44	7.33	9.92	3.00	4.65	12.33	19.09	6.00	5.40
Other students	11.25	15.04	3.13	2.13	1.48	1.41	6.04	8.00	10.84	5.48	2.85	1.33	13.11	13.01	7.45	2.29

Note. Abbreviations: S1 = first semester and S2 = second semester.

[a] Italicized measures are from interaction during reading groups. Nonitalicized measures are from general class activities.

failure situations, three showed radiation effects such that they also tended to stay with other students longer. The other four teachers showed large gains for the extension students and little change for their classmates. The teacher that did not change her behavior toward extension students did not change her behavior toward classmates, either.

In summary, the treatment for the extension students was effective with seven of the eight teachers, and gains for extension students were achieved without loss to their classmates. Further, in three classes, the gains for extension students radiated so as to benefit their classmates also.

Analyses of variables less central to the treatment of extension students showed that teachers tended to ask them more questions during reading groups following treatment. This increase was not at the expense of classmates except in one classroom where the reading group response opportunities were more than doubled for extension students while being cut in half for their classmates. In summary, the effects of the treatment for the extension students were rather general across teachers but were confined mostly to the measures of teacher behavior in staying with students following failure. For the most part, the advantages accruing to the extension students as a result of the treatment were not gained at the expense of classmates (although there was one exception), and the treatment sometimes radiated to the benefit of classmates.

The treatment regarding *low-participation* students produced large gains in the frequencies of response opportunities and interactions afforded by the teachers. In a sense, these quantitative gains were at the expense of classmates, since the means for classmates went down in most classes, while the means for the low-participation students went up. However, the effect of the treatment was to more nearly equalize response opportunities and teacher–student contacts for low-participation students and their classmates rather than to make the teachers spend most of their time with the low-participation students and ignore their classmates. Further, as noted previously, even after improvement following treatment, most of the measures of frequency of contacts with teachers showed low-participation students to be still behind their classmates in the second semester.

Discussion

Clearly, the first treatment goal—changing specified teacher behavior with target children—was accomplished. Teachers substantially increased the number of contacts they had with low participants, and they also increased the percentage of times they continued to work with extension students after they responded inadequately. Thus, the consultation strategy of making teachers aware of their differential behavior toward students in quasi-equivalent circumstances was effective in changing teacher behavior.

Teachers changed their behavior more toward low-participation students, even though during the treatment interview they expressed more concern about and interest in changing their behavior toward extension students. This suggests that the basic hypothesis, inappropriate teacher behavior is due simply to unawareness, might be more appropriate to the low-participation group than to the extension group. The data from the low-participation group suggest that the teachers were broken out of a rut by our treatment: the treatment made them see these students in a whole new way, and it also made the students respond somewhat differently than they had in the past. In contrast, while the treatment succeeded in getting the teachers to stay with extension group students longer, it did not remove some of the other negative treatment of these students and, in general, did not lead to such widespread or beneficial effects as the treatment with low participants. This may mean that teacher behavior toward extension students is controlled by student behavior and subtle affective teacher reaction to this behavior.

Perhaps a more systematic feedback plan would have produced greater changes; however, the data show that a relatively simple and brief treatment was sufficient to change the teacher behaviors specified in this

study. The basic strategy—identifying and reinforcing certain desirable teaching behaviors and requesting that teachers transfer them to additional students—appears potentially applicable for producing a wide variety of desirable changes in school settings.

It is important to note that the investigators alerted the teachers to their differential treatment of different students[3] but only *suggested* that target students might respond more favorably if they were treated in the same way as contrast children. Some time was spent suggesting how teachers could change their interaction patterns with target children, but few specific techniques were prescribed. It was left to the teachers to decide whether and how to operationalize these principles (increase contact, stay with the student). It should also be noted that the treatment involved bringing in the teacher as a partner rather than attempting to manipulate her through deception or provision of phony information. We think that this was partly responsible for the treatment's success as was the fact that all teachers were enthusiastic toward the study and perceived it as helpful and relevant. Thus, there is much to be gained from research which makes the investigator a partner and a resource person to the teacher rather than that which makes him a manipulator of the teacher.

The effects of the treatment on student behavior were less clear than the effects on teacher behavior. There were signs that the low participants began seeking out the teacher more than they had previously, although the gains were relatively small. There was also slight evidence to show that extension students changed. Thus, both groups were influenced in some ways by the treatment program. Although the changes were small, the students did appear to react reciprocally to improved teacher behavior.

The effects of the treatment program on nontarget children are of great experimental and practical interest. The value of changing teacher behavior toward two or three children is dubious if teacher interaction patterns with other children deteriorate. However, teacher changes toward target children typically did not interfere with other teacher–child interaction patterns. In fact, when teachers did change their behavior toward target children, they also tended to change their behavior (in the same direction) toward nontarget children, at least on measures of qualitative aspects of teacher–student interaction.

Another question of methodological interest is the possible contamination effect (either positive or negative) of the treatment upon teacher interaction with target children. Teachers may change their behavior toward target children in ways *not* specified by the treatment (teachers asked to "stay with" may also ask more frequent and more difficult questions). The nature and desirability of such contamination effects seem worthy of experimental investigation. Such effects may or may not appear in a given treatment; however, an investigator cannot discover or study them unless he includes a wider variety of interaction variables than those involved in the treatment.

Contamination effects were operative in this study and most, but not all, were positive. However, there were some negative effects. For example, teachers criticized extension students, when they gave a wrong answer or when they made no response during reading group, more often than they had previously. Perhaps the teachers were doing the wrong thing for the right reason. That is, the teachers were determined to work with these students and to demand academic responses; perhaps when their responses were inadequate, teacher disappointment spilled out in the form of criticism. In any case, it seems necessary that future investigators, especially if they make complex treatment requests, monitor

[3] This is not to suggest that teachers should treat all children in the same way. The point that students need different patterns of classroom contact with their teacher has been discussed at length elsewhere (Brophy & Good, 1974). In this paper, the term "treat the same way" is used to describe the idea that was presented to teachers: "You have already demonstrated that you can perform the behavior with similar students under similar circumstances; provide it for these students as well."

teacher behavior for possible inappropriate changes. Treatments that make students more salient to their teachers may produce undesirable side effects.

Finally, note that for both groups, teachers focused more on the academic behavior of the students as a result of the treatment. In both cases, the teachers worked for improved performance. However, they appeared to push and to challenge the extension students while being much more supportive and protective with low participants.

In summary, the study has demonstrated that a simple consultation strategy for presenting teachers with feedback about their behavior was effective in changing both quantitative and qualitative teacher behavior toward target students. In addition, student behavior was influenced by the change in teacher behavior. The effect of the treatment on nontarget children was not detrimental. Of special interest is that specified or requested change in teacher behavior toward target children is likely to be accompanied by additional changes in teacher behavior toward target and nontarget children. Predicting and controlling such effects should be of special concern to those who propose to change teacher behavior. Finally, a call has been made for more research which involves the classroom teacher as a research partner rather than as a subject to be manipulated.

REFERENCES

BROPHY, J., & GOOD, T. Brophy-Good system (teacher-child dyadic interaction). In A. Simon and E. Boyer (Eds.), *Mirrors for behavior: An anthology of observation instruments continued, 1970 supplement*. Vol. A. Philadelphia, Penn. Research for Better Schools, 1970. (a)

BROPHY, J., & GOOD, T. Teachers' communications of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology*, 1970, **61**, 365–374. (b)

BROPHY, J., & GOOD, T. *Teacher-student relationships: Causes and consequences*. New York: Holt, Rinehart & Winston, 1974.

DEGROAT, A., & THOMPSON, G. A study of the distribution of teacher approval and disapproval among sixth-grade pupils. *Journal of Experimental Education*, 1949, **18**, 57–75.

GOOD, T. Which pupils do teachers call on? *Elementary School Journal*, 1970, **70**, 190–198.

GOOD, T., & BROPHY, J. Teacher-child dyadic interactions: A new method of classroom observation. *Journal of School Psychology*, 1970, **8**, 131–138.

GOOD, T., & BROPHY, J. Analyzing classroom interaction: A more powerful alternative. *Educational Technology*, 1971, **11**, 36–41.

GOOD, T., & BROPHY, J. Behavioral expression of teacher attitudes. *Journal of Educational Psychology*, 1972, **63**, 617–627.

GOOD, T., & BROPHY, J. *Looking in classrooms*. New York: Harper & Row, 1973.

HORN, E. *Distribution of opportunity for participation among the various pupils in classroom recitations*. New York: Columbia University, Teachers' College, 1914.

JACKSON, P., & LAHADERNE, H. Inequalities of teacher-pupil contacts. *Psychology in the Schools*, 1967, **4**, 204–208.

JONES, V. *The influence of teacher-student introversion, achievement, and similarity on teacher-student dyadic classroom interactions*. Unpublished doctoral dissertation, University of Texas at Austin, 1971.

KRANZ, P., WEBER, W., & FISHELL, K. The relationships between teacher perception of pupils and teacher behavior toward those pupils. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March 1970.

MCNEIL, J. *Toward accountable teachers: Their appraisal and improvement*. New York: Holt, Rinehart & Winston, 1971.

MENDOZA, S., GOOD, T., & BROPHY, J. *Who talks in junior high classrooms?* (Report Series 68) Austin: University of Texas, Research and Development Center for Teacher Education, 1972.

RIST, R. Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard Educational Review*, 1970, **40**, 411–451.

ROWE, M. Science, silence, and sanctions. *Science and Children*, 1969, **6**, 11–13.

SIKES, J. *Differential behavior of male and female teachers with male and female students*. Unpublished doctoral dissertation, University of Texas at Austin, 1971.

SILBERMAN, M. Behavioral expression of teachers' attitudes toward elementary school students. *Journal of Educational Psychology*, 1969, **60**, 402–407.

TUCKMAN, B., & OLIVER, W. Effectiveness of feedback to teachers as a function of source. *Journal of Educational Psychology*, 1968, **59**, 297–301.

TURNER, J., FOA, E., & FOA, U. Interpersonal reinforcers: Classification, interrelationship and some differential properties. *Journal of Personality and Social Psychology*, 1971, **19**, 168–180.

WITHALL, J. An objective measurement of a teacher's classroom interaction. *Journal of Educational Psychology*, 1956, **47**, 203–212.

(Received July 23, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 406-409

THE EFFECTS OF READABILITY ON ORAL AND SILENT READING RATES[1]

ESTHER U. COKE[2]

Bell Laboratories, Murray Hill, New Jersey

The usefulness of reading rate as a measure of reading difficulty was evaluated in two studies relating reading rate to two text-derived measures of readability. In the first study, oral reading rate was unaffected by readability since the subjects read at a constant syllable rate. A second study attempted to increase the saliency of the readability variable by having the subjects rate each passage for comprehensibility. Passages were read both silently and aloud at a constant syllable rate. Comprehensibility ratings were correlated with readability indicating that the subjects were sensitive to readability. These results suggested that there are important limitations on the usefulness of reading rate as a measure of reading difficulty.

Recent experimental evidence has challenged the validity of the assumption that reading rate is a good behavioral index of the readability of English prose. When reading rate is measured in units smaller than a word, the rate remains constant over a wide range of difficulty (Carver, 1971; Miller & Coleman, 1971; Sticht, 1971). The present experiment is a replication of the Miller and Coleman study, using a much larger sample of texts and more precise experimental techniques.

Reading rate is a potentially useful measure in studies of language processing since rate can be easily derived from observations of reading. Therefore, it is important to determine the sensitivity of this measure to factors that influence the difficulty of processing written materials.

Experiment 1

Method

Material. Ninety prose passages differing in subject matter and difficulty were selected from the following sources: 29 passages from the 36 used by Miller and Coleman (1971); 28 passages from the 330 used by Bormuth (1969); 12 selections from a high school physics text; 14 selections from randomly chosen professional journals in chemistry, electronics, and economics; and 7 selections from randomly chosen nontechnical books in a university library. These passages covered the following subject matters: narratives and popular science (22 texts); history, civics, and literature (17 texts); economics (6 texts); biology and psychology (18 texts); and chemistry, physics, and electronics (27 texts). Five additional passages were selected from the same sources and were used to control for warm-up effects.

Computer-generated negative microfilm prints were made of all passages except the Miller-Coleman selections. These prints were mounted in slides. The Miller-Coleman passages had been typed and photographed, and the positive prints had been mounted in slides for another study. The computer-generated letters were slightly larger than the typewritten letters.

Procedure. Each subject read all passages during two reading sessions held on separate days. At the beginning of the first session, the subject read experimental directions which emphasized the need to read rapidly without sacrificing intelligibility. The subject was told that his reading would be recorded to check for clarity and that his reading times would be clocked. The subject then read aloud 3 of the warm-up passages and 26 of the experimental passages. After a five-minute rest, 25 more experimental passages were read aloud. At the beginning of the second day's session, the experimenter reviewed the oral reading directions. Then the subject read aloud 2 of the warm-up passages and 14 experimental passages. After a five-minute rest, 25 more experimental passages were read aloud.

Over both sessions, all the subjects read the

[1] The author wishes to thank E. Z. Rothkopf for his advice and suggestions. Portions of this paper are based on a paper read at the meeting of the American Educational Research Association, New Orleans, February 1973.

[2] Requests for reprints should be sent to Esther U. Coke, Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974.

same 5 warm-up passages in the same order. Each subject read the 90 experimental passages in a different random order.

Apparatus. The subject controlled the exposure time of all slides with a switch that also activated a timer. Reading time was recorded in tenths of a second. Each passage slide was followed by a blank slide so that the subject could rest after reading aloud. The slides were shown on a rear-projection screen.

Subjects. Twenty Montclair State College undergraduates served as paid volunteers for the study.

Text measures. Two measures of length were obtained for each passage: the total number of words and an estimate of the total number of syllables. The syllable count was estimated from the number of vowels according to a formula used by Coke and Rothkopf (1970). Two indices of readability were calculated for each passage: the average word length in syllables and the average sentence length in words. The simplified Flesch reading ease score (Flesch, 1949) was also calculated. This score is the weighted sum of the average word length in syllables and the average sentence length in words.

Results

To determine if the reading times for the 29 Miller-Coleman passages differed from the times for the other 61 passages because of differences in typography, a one-way analysis of covariance was used with the total number of words, the average word length in syllables, and the average sentence length in words as covariates. The mean reading times of the two sets of passages did not differ significantly ($F = .03$, $df = 85, p > .05$). Therefore, all 90 passages were used in the analysis of the data.

Reading time and passage length. The mean reading time for each passage was calculated over all 20 subjects and correlated with both measures of passage length. Number of syllables was a better predictor of reading time ($r = .94$) than was number of words ($r = .76$). The difference between these correlations was statistically significant when the correlation between syllable and word length ($r = .72$) was taken into account. The large correlation between syllable length and reading time implies that subjects read all passages at a reasonably constant syllable rate.

Reading rate and readability. Figure 1 clearly shows that when reading rate was measured in syllable units, rate remained constant over a wide range of difficulty. When the two indices of readability were considered separately, neither average word length in syllables nor average sentence length in words was a good predictor of syllable rate. Average word length in syllables accounted for 12% ($r = .34$) of the variability in rate while average sentence length in words accounted for less than 1% ($r = -.05$) of the variability.

Figure 1 also shows that when reading rate was measured in word units, hard passages were read more slowly than easy passages. The large correlation ($r = .90$) between difficulty and word rate can be attributed to differences in average word length in syllables. Average word length in syllables accounted for 83% ($r = -.91$) of the variability in word rate while average sentence length in words accounted for only 26% ($r = -.51$) of the variability.

Discussion

The results clearly support Miller and Coleman's (1971) conclusion that reading rate is constant over a wide range of difficulty when rate is measured in units smaller than a word. The finding that word rate decreased with difficulty can be accounted for by the observed syllable rate constancy. Since subjects read at a constant syllable rate, words containing more syllables took longer to read than words with fewer syllables. Therefore, the harder passages, which had a larger proportion of longer words, were read more slowly when rate was measured in word units.

Two explanations of the observed reading rate constancy suggest themselves. One is that the output requirements of speaking make oral reading rate insensitive to passage difficulty. Another explanation is that the reader does not need to understand a passage when he reads it aloud. In order to test the adequacy of these explanations, a second experiment was conducted. The reading task in this experiment involved comprehension and allowed for an independent evaluation of the reader's sensitivity to passage difficulty. In addition, the passages were read both aloud and silently.

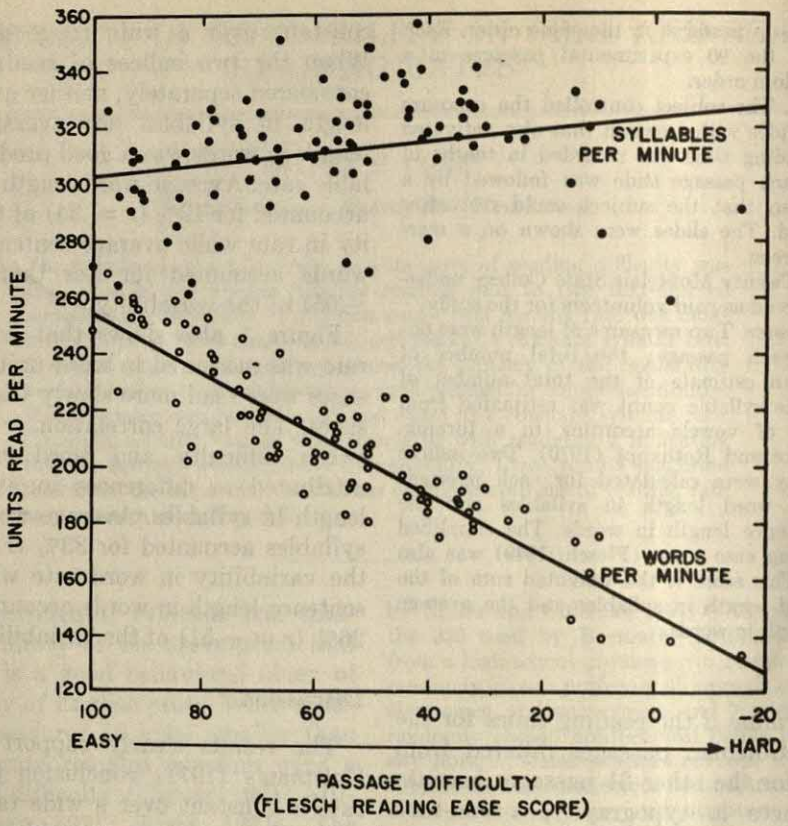


FIGURE 1. Scatter plots showing reading rate in syllables per minute (filled dots) and in words per minute (open dots) as a function of passage difficulty. (Straight lines were fitted to the data by the method of least squares.)

EXPERIMENT 2

Method

Material. Thirty-two passages were selected from the 90 passages of Experiment 1. These 32 passages covered the same range of length and difficulty as the larger set. Miller-Coleman (1971) passages were generated on microfilm by computer so that the passages did not differ in typography.

Procedure. Each subject read all 32 passages under one of three reading conditions. One group (aloud–pronounce) read each passage aloud and judged how difficult the words were to pronounce; a second group (aloud–understand) read each passage aloud and judged how difficult the passage was to understand; a third group (silent–understand) read each passage silently and also rated the passage for comprehension difficulty.

Each passage was rated on a 5-point scale ranging from very hard (1) to very easy (5). Passage slides alternated with blank slides. Ratings were recorded when the blank slide was on the screen.

A short-answer comprehension test was given immediately after the reading session. Subjects were unaware that they would be tested. There were two questions about each passage. A control group (no reading) of 26 subjects took the test without reading the passages first.

Reading directions were the same as in Experiment 1. Silent readers were urged to read each passage only once from the beginning to the end. All subjects read 2 warm-up passages followed by the 32 passages in random order.

Apparatus. The same apparatus was used as in Experiment 1.

Subjects. Eighty-six paid volunteer high school students participated in the study. Twenty subjects were assigned to each of the three reading conditions.

Text measures. The measures of passage length and readability were the same as those in Experiment 1.

Results

Reading time and passage length. The results for Experiment 1 were replicated. Under all reading conditions, subjects read the passages at a reasonably constant syllable rate (r = .99, .98, and .94, respec-

tively, for conditions aloud–pronounce, aloud–understand, and silent–understand).

Reading rate and readability. As in Experiment 1, reading rate in syllables was constant over the entire range of difficulty regardless of reading condition. Requiring subjects to evaluate comprehensibility did not make oral reading rate any more sensitive to readability than it was when subjects made pronounceability ratings or simply read aloud as in the first experiment. Further, the removal of the output constraint of speaking for the silent readers did not make syllable rate more sensitive to the readability indices.

Comprehensibility and readability. Multiple correlations were calculated between the average comprehensibility ratings of the passages and the two readability indices—average word length in syllables and average sentence length in words. Jointly, these two indices accounted for 56% of the variability in the ratings of oral readers (r = .75) and 58% of the variability of silent readers (r = .76). The results show that subjects were responsive to text features associated with readability.

In addition, subjects in all conditions remembered a good deal of the passages' content. Subjects who read the passages answered over 50% of the postreading questions correctly. Subjects who had never read the passages before (no-reading condition) answered less than 1% of the questions correctly.

These findings rule out an explanation of the observed syllable rate constancy in terms of superficial processing. Subjects in this experiment were not simply pronouncing words without comprehension when reading aloud or were not simply glancing at the texts when reading silently.

Discussion

The results of this study fully support the contention that both oral and silent reading rate remain constant across the entire range of difficulty of English prose when rate is measured in a unit smaller than a word (Miller & Coleman, 1971). The almost universal practice of measuring rate in words can lead to spurious conclusions about the relationship between reading rate and readability. In this experiment, the large positive correlation between word rate and passage difficulty could be explained in terms of the constancy of reading rate in syllables. Educational researchers would be prudent to look at syllable rate when assessing the effects of readability on reading rate.

While subjects read at a constant syllable rate in this experiment, it is entirely possible that changes in instructions or other task variables might have caused the readers to adapt their rate to passage difficulty (Kershner, 1964; Miller & Coleman, 1971). This dependence of reading rate on set factors places important limitations on the usefulness of both oral and silent reading rate in studies of reading.

REFERENCES

Bormuth, J. C. Development of readability analyses. (U.S. Department of Health, Education, and Welfare Final Report, Project 7-0052, Contract EC-3-7-070052 0326) Chicago: University of Chicago, 1969.

Carver, R. P. Evidence for the invalidity of the Miller-Coleman readability scale. *Journal of Reading Behavior*, 1971, **4**, 42–47.

Coke, E. U., & Rothkopf, E. Z. Note on a simple algorithm for a computer-produced reading ease score. *Journal of Applied Psychology*, 1970, **54**, 208–210.

Flesch, R. *The art of readable writing*. New York: Harper, 1949.

Kershner, A. M. Speed of reading in an adult population under differential conditions. *Journal of Applied Psychology*, 1964, **48**, 25–28.

Miller, G. R., & Coleman, E. B. The measurement of reading speed and the obligation to generalize to a population of reading materials. *Journal of Reading Behavior*, 1971, **4**, 48–56.

Sticht, T. G. *Learning by listening in relation to aptitude, reading, and rate-controlled speech: Additional studies*. (HumRRO Tech. Rep. 71-5) Alexandria, Va.: Human Resources Research Organization, 1971.

(Received July 30, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 410-415

A STATISTICAL TEST OF THE THEORETICAL MODEL FOR THE REPRESENTATIONAL LEVEL OF THE ILLINOIS TEST OF PSYCHOLINGUISTIC ABILITY[1]

EUGENE B. DOUGHTIE,[2] JAMES A. WAKEFIELD, JR., RAYMOND N. SAMPSON, AND HERBERT L. ALSTON

University of Houston

The correspondence between the relative placement of the first six subtests of the Illinois Test of Psycholinguistic Ability (ITPA) in factor space with their theoretical placement in the representational level of the ITPA model was investigated using data from the eight age levels presented in Paraskevopoulos and Kirk. The relative placement of the six subtests in factor space was unrelated to their positions in the model for the 3- and 4-year-old children. The placement of the subtests showed significant correspondences with the model for each of the six older groups—age 5 through 10.

The Illinois Test of Psycholinguistic Abilities (ITPA; Kirk, McCarthy, & Kirk, 1968; McCarthy & Kirk, 1961) is an instrument for assessing the language abilities of children. The revised test is composed of 12 subtests designed to measure a three-dimensional model of communication (clinical model). The test is conceptualized in terms of channels of communication, processes, and levels of organization. The two channels incorporated in the test are the auditory-vocal and the visual-motor. The processes refer to reception, organization, and expression. The two levels of organization are the representational level and the automatic level. Items in the 6 representational-level subtests are thought to require more abstract reasoning abilities, whereas the items at the automatic level are thought to be responded to in a more habitual manner. Each of the 6 subtests at the representational level is designed in reference to a channel and a process. The automatic-level subtests are not considered to be as amenable to measuring separate processes. An attempt is made, however, to consider a single channel, level, and process with each of the subtests.

The ITPA model is a modification of a theory presented by Osgood (1957a, 1957b). The authors of the ITPA utilized 12 of the 48 possible combinations of channels, levels, and processes suggested by the Osgood model (Carroll, 1972). The representational-level subtests are Auditory Reception, Visual Reception, Auditory Association, Visual Association, Verbal Expression, and Manual Expression. The automatic-level subtests are Visual Memory, Auditory Memory, Visual Closure, Grammatic Closure, Auditory Closure, and Sound Blending. A more complete discussion of the subtests may be found in the *Examiner's Manual* (Kirk et al., 1968).

There has been little support for this structure either experimentally or from factor-analytic studies of the ITPA (Dale, 1972). The major emphases of the factor-analytic studies of the ITPA have been to determine whether the subtests of the ITPA actually measure several different abilities as has been claimed (Kirk & Kirk, 1971) and to determine whether factors obtained

[1] Samuel A. Kirk generously agreed to allow us to use data from Tables A-1 through A-8 (Paraskevopoulos & Kirk, 1969) that made this study possible.

[2] Requests for reprints should be sent to Eugene B. Doughtie, Department of Psychology, Cullen Boulevard, University of Houston, Houston, Texas 77004.

lend support to the clinical model of the ITPA.

Factor-analytic studies of the experimental edition of the ITPA (McCarthy & Kirk, 1961) have typically reported three or four factors (Carroll, 1972; Meyers, 1969; Silverstein, 1967). Differences in the factors at different ages have been reported (Ryckman & Wiegerink, 1969; Silverstein, 1967); more factors have been obtained with older children than with younger children (Ryckman & Wiegerink, 1969); and social class appears to affect the factors obtained (Uhl & Nurss, 1970). It has been noted that the use of reference tests to extend the variance in the correlation matrix may increase the number of factors obtained and facilitate the interpretation of the factors (Carroll, 1972; Meyers, 1969; Ryckman & Wiegerink, 1969).

Burns and Watson (1973) factor analyzed the 12 subtests of the revised ITPA (Kirk et al., 1968), using a group of underachieving children as subjects, and obtained five factors. They interpreted the factors as supporting existence of the visual-motor and auditory-vocal channels of the clinical model of the ITPA as well as supporting that of the process of expression. However, support for the reception and association processes and differentiation between the automatic and representational levels of the theoretical arrangement was not found.

Two explanations for the lack of definitive support for the arrangement present themselves. The first is, of course, that the ITPA subtests may not yield data that conform to the structure. The second possibility which must be considered is that, although the ITPA may correspond to the model in some degree, the factor-analytic studies of the test scores have failed to find the correspondence. This might occur in two ways.

First, the factor patterns obtained by analytic methods of rotation (c.f. Harman, 1970) may have failed to correspond to the three dimensions of the configuration closely enough to be recognized by visual inspection. This seems especially likely with a structure, such as the ITPA model, that predicts a uniform distribution of points in factor space rather than a number of clusters of points. In particular, the varimax method of rotation maximizes the variance of a factor by rotating it toward points relatively close together. A slight displacement of one point in the direction of another nearby point would tend to "attract" a factor, possibly causing the rotated factor to miss a theoretically more meaningful position. The second possibility is that the three axes of the spatial configuration may not be measured orthogonally by the ITPA. If the three axes are correlated, an orthogonal factor analysis might find one or two of the dimensions of the structure and fail to indicate the others.

Several of the studies mentioned above (Burns & Watson, 1973; Meyers, 1969; Ryckman & Wiegerink, 1969) have found partial although not definitive support for the clinical model of the ITPA. The present study deals directly with the correspondence between the ITPA and the clinical model underlying its development and interpretation by using a factor-analytic method that is unaffected by the idiosyncracies of rotation and which should be tolerant of a moderate degree of correlation of the axes of the arrangement. This method has been recently employed to test the correspondence between a personality inventory and its underlying model (Wakefield & Doughtie, 1973). As there is reason to believe that the factors of the ITPA change with age, the correspondence is tested independently for the eight age groups presented in Paraskevopoulos and Kirk (1969).

Method

Analysis of the Model in Terms of Distance Comparisons

In order to test the correspondence of the subtests of the ITPA with the clinical model, it is necessary to consider the relationships between pairs of subtests in terms of distances between the corresponding points in the spatial configuration. The order relationships among the distances implied by the structure will then serve to define the model for the purpose of the test.

The conceptualization, as presented by Paraskevopoulos and Kirk (1969), implies order relationships among the distances between the points

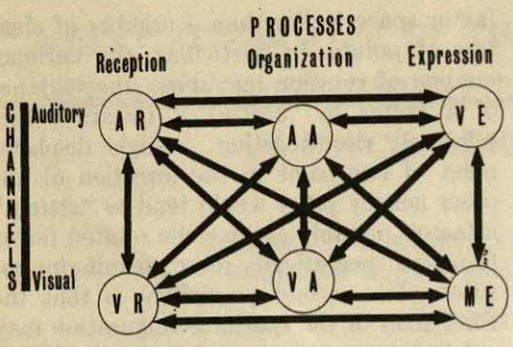


FIGURE 1. Representational level of the Illinois Test of Psycholinguistic Ability model. (Abbreviations: AR = Auditory Visual, VR = Visual Reception, AA = Auditory Association, VA = Visual Association, VE = Verbal Expression, ME = Manual Expression.)

in the representational level of the model. Unfortunately, they were not specific about the positions of the automatic, "whole level" subtests. As a result, no hypotheses about distances involving these points were made. However, the positions of the six subtests at the representational level were clearly indicated. These six subtests were located at the intersections of the two channels of communication—auditory and visual—with the three processes—reception, organization, and expression. The representational level of the model is presented in Figure 1.

By constructing a complete graph of these six points (i.e., connecting every point with every other point; c.f. Coombs, Dawes, & Tversky, 1970, pp. 78–94), all the interpoint distances in the theoretical arrangement are represented as lines connecting two points. A complete graph of six points contains 15 lines. The rectangular structure of the representational-level subtests clearly implies that certain of these lines should be longer than others.

Seven of the lines in Figure 1 are *short sides*. These lines connect points that are in the same channel and adjacent processes or that are in the same process and different channels. Four lines are *short diagonals* connecting points in different channels and adjacent processes. Two lines are *long sides* connecting points that are in the same channel and in nonadjacent processes. The remaining two lines are *long diagonals* connecting points that are in different channels and nonadjacent processes.

The following order relations of interpoint distances follows from the configuration. The two *long diagonals* are longer than all the other lines. The seven *short sides* are shorter than all the other lines. The *short diagonals* and the *long sides* are intermediate in length.

The lack of specificity in the literature on the ITPA concerning the relative separation among processes and between channels requires that two sets of order relationships be given special attention. The first set is the comparisons of the two *long sides* with the three *short sides* that connect points in different channels. Depending on the relative separation along the two dimensions, the three short distances could be classified as short or long. They were classified as short for two reasons. The ITPA manual (Kirk et al., 1968) depicted the two channels as closer than the reception and expression processes. A methodological reason (Campbell & Fiske, 1959) is that the channels may be considered different methods of measuring the three theoretical constructs—reception, organization and expression. As the two methods may be different due to artifacts of measurement, placing the two measures of each construct relatively close together demonstrates the validity of the three processes.

The second set of order relationships due special attention is the possible comparisons between the *short diagonals* and the *long sides*. Following the reasoning of the previous paragraph, it would be desirable for the short diagonals to be shorter than the long sides. However, the relationship between these two sets of distances is influenced by factors other than the relative separation of points along the two dimensions. The most important influences are the angles between the sides. A slight deviation from orthogonality would change the lengths of the diagonals without changing the lengths of the sides. As lack of orthogonality of the dimensions was considered a possible reason for the inconclusive results of previous factor-analytic study of the ITPA, comparisons so largely influenced by variations in the angles separating the sides were avoided. Hence, comparisons of *long sides* and *short diagonals* were not employed in the test of the model.

The 15 interpoint distances are listed in Table 1 as members of three classes defined by their relative lengths as derived from the model. The classes are called the short, middle, and long classes. Each of the seven short lines is shorter than each of the six middle lines; this yields 42 pairs of ordered distances. Each of the six middle lines is shorter than each of the two long lines; this yields 12 pairs of ordered distances. In total, 54 pairs of ordered distances describe the representational level of the model.

Data

Eight matrices of the intercorrelations of the 12 subtests of the revised ITPA from Paraskevopoulos and Kirk (1969, pp. 202–209) for eight different ages were used to test the model. The first matrix was obtained on a sample of 107 children aged 2 years 7 months to 3 years 1 month; the second, on a sample of 116 children aged 3 years 7 months to 4 years 1 month; the third, on a sample of 115 children aged 4 years 7 months to 5 years 1 month; the fourth, on a sample of 128 children aged 5 years 7 months to 6 years 1 month;

the fifth, on a sample of 124 children aged 6 years 7 months to 7 years 1 month; the sixth, on a sample of 123 children aged 7 years 7 months to 8 years 1 month; the seventh, on a sample of 127 children aged 8 years 7 months to 9 years 1 month; the eighth, on a sample of 122 children aged 9 years 7 months to 10 years 1 month.

Factor Analyses

In order to test the correspondence between the empirical placement of subtests in factor space and their theoretical arrangement against a null hypothesis of random placement, it was necessary to consider the dimensionality of the factor space. Shepard (1962a, 1962b) has shown that the distances among a set of N points can take any conceivable rank order in a space of at least $N - 1$ dimensions. In order to allow the possibility of the occurrence of any rank order of the distances among the six representational-level subtests, a space of at least five dimensions was necessary. Thus, the first five principal components were extracted from each of the eight matrices. The five components were rotated by the varimax method (Harman, 1970).

The eight analyses accounted for approximately the same percentages of total variance at each age, ranging from 61.1% to 67.9%.

Distances in Factor Space

Euclidean distances from every point to every other point in factor space were computed by a formula representing a generalization of the familiar Pythagorean theorem to multidimensional space (c.f. Wakefield & Doughtie, 1973).

TABLE 1

FIFTY-FOUR ORDER RELATIONS OF DISTANCES DERIVED FROM THE ITPA MODEL

42 ordered pairs			12 ordered pairs		
Short distances	are less than	Middle distances	Middle distances	are less than	Long distances
AR–VR		VA–AR	VA–AR		
AA–AR		AA–VR	AA–VR		ME–AR
VA–VR		ME–AA	ME–AA		VE–VR
VA–AA	<	VE–VA	VE–VA	<	
VE–AA		VE–AR	VE–AR		
ME–VE		ME–VR	ME–VR		
ME–VA					

Note. Abbreviations: ITPA = Illinois Test of Psycholinguistic Ability, AR = Auditory Reception, VR = Visual Reception, AA = Auditory Association, VA = Visual Association, VE = Verbal Expression, and ME = Manual Expression.

TABLE 2

PAIRS OF DISTANCES IN FACTOR SPACE CORRESPONDING TO THE ITPA MODEL AT EIGHT AGES

Age (years-months)	Correct orders (out of 54)	Significance level
2–7 to 3–1	22	*ns*
3–7 to 4–1	26	*ns*
4–7 to 5–1	46	.01
5–7 to 6–1	39	.01
6–7 to 7–1	37	.01
7–7 to 8–1	35	.05
8–7 to 9–1	42	.01
9–7 to 10–1	36	.05

Note. Abbreviation: ITPA = Illinois Test of Psycholinguistic Ability.

Comparing Orders of Obtained Distances to the Model at Each Age

The order of each pair of distances in factor space for which the ITPA model yielded a specific prediction was obtained for each age group. The total number of empirical distances that showed the correct theoretical order was obtained for each age.

Significance of the Correspondence of the ITPA to the Model at Each Age

In a space of five dimensions, any distance between two of six points is equally likely to be greater than or less than any other distance between two of the six points (Shepard, 1962a, 1962b). Hence, each of the 54 comparisons of a theoretical ordering of two distances with the corresponding empirical order represents a Bernoulli trial (with $p = .5$).

Using the normal approximation to a binomial test for goodness of fit (Siegel, 1956, pp. 36–42), it was found that 34 comparisons between two distances occurring in the predicted order were necessary to reject the null hypothesis of random distance orders at the .05 level and 37 were necessary to reject at the .01 level. Rejection of the null hypothesis would, of course, support the contention that the ITPA subtests are interrelated in the fashion prescribed by the model.

RESULTS

Table A contains the eight factor patterns, the eight sets of interpoint distances in factor space, and the eight lists of correct and incorrect distance pair orders.[3]

[3] A 10-page table containing the eight factor patterns, the eight sets of interpoint distances in

Correspondence to the Model

Table 2 presents the number of pairs of distances that had the same order as that predicted from the model for each age level. For the two youngest levels, there seemed to be no relationship between the ITPA subtests and the representational level of the model. The number of correct orders by chance alone is 27 (half of the 54 observations). The orders for both of the two youngest ages are close to (in fact, slightly below) chance level. The subtests of the ITPA did show significant correspondences to the representational level of the model for the six older groups of subjects. Four ages yielded significance at the .01 level and two at the .05 level.

Correspondence to the Model and Age of Subjects

A Spearman rank-order correlation (Siegel, 1956, pp. 202–213) between the rank order of age and the rank order of the number of correct distance comparisons was computed to find evidence of a developmental trend toward increasing correspondence to the theoretical structure with increasing age. The correlation ($r = .36$) was not significant ($p > .05$).

Discussion

The clinical model of the ITPA was reliably approximated by the representational-level subtests on data obtained from the six oldest age groups treated by Paraskevopoulos and Kirk (1969) but not on data obtained from the two youngest age groups. The age groups that ranged from 2 years 7 months to 3 years 1 month and from 3 years 7 months to 4 years 1 month yielded patterns of the six representational-level subtests which had no demonstrable relationship to the clinical model. The children in the six older groups, near their fifth, sixth, seventh, eighth, ninth, or tenth birthdays, yielded subtest patterns that approximated the pattern of the representational level of the clinical model.

An inspection of Table 2 suggests that the groups did not gradually approach the theoretical arrangement with advancing age, as an incremental view of language development would predict. Rather, the change from no relationship to the ITPA conceptualization to that of a significant correspondence to the model occurred within a single year. There are, of course, fluctuations in the degree to which each age approximates the configuration, but no relationship between correspondence to the structure and the age of the subjects is suggested except the one large change between the group of children who were approximately four years old and the group who were approximately five.

These results have bearing on the appropriateness of some interpretations drawn from ITPA score patterns for children at different ages. While interpretations based on empirical prediction of a criterion are not affected by the appropriateness or inappropriateness of the clinical model, more theoretical interpretations based on the theoretical arrangement must be held suspect for children who are four or younger. Interpretations of the ITPA based on the model may, however, be reasonably applied to children who are approaching or have passed their fifth birthday.

An important milestone in the development of children is suggested by these results. Between the ages of 2 and 10 years, there is apparently a single, rather sudden, change from a lack of correspondence to the model of communication at the representational level to an enduring correspondence to the theoretical structure. The age at which the change occurs corresponds roughly to the end of Piaget's preconceptual phase (of the preoperational stage) and the beginning of the intuitive phase (Longstreth, 1968). The important distinction between children at these levels is that the child in the preconceptual phase is not able to perform operations concerning classes, relationships, and number, while the intuitive child can perform such operations, although not as parts of larger systems of operations. A closer look at the items comprising the subtests of the ITPA at the representational level may reveal many items

factor space, and the eight lists of correct and incorrect distance pair orders is available from the senior author upon request.

that involve cognitive operations that children younger than about 4 years of age cannot perform.

REFERENCES

BURNS, G. W., & WATSON, B. L. Factor analysis of the revised ITPA with underachieving children. *Journal of Learning Disabilities*, 1973, **6**, 371–376.

CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, **56**, 81–105.

CARROLL, J. B. Review of ITPA. In O. K. Buros (Ed.), *The seventh mental measurements yearbook*. Highland Park, N. J.: Gryphon Press, 1972.

COOMBS, C. H., DAWES, R. M., & TVERSKY, A. *Mathematical psychology*. Englewood Cliffs, N. J.: Prentice-Hall, 1970.

DALE, P. S. *Language development: Structure and function*. Hinsdale, Ill.: Dryden Press, 1972.

HARMAN, H. H. *Modern factor analysis*. Chicago: University of Chicago Press, 1970.

KIRK, S. A., & KIRK, W. D. *Psycholinguistic learning disabilities: Diagnosis and remediation*. Urbana: University of Illinois Press, 1971.

KIRK, S. A., MCCARTHY, J. J., & KIRK, W. D. *Examiner's manual: Illinois Test of Psycholinguistic Abilities*. Urbana: University of Illinois Press, 1968.

LONGSTRETH, L. E. *Psychological development of the child*. New York: Ronald Press, 1968.

MCCARTHY, J. J., & KIRK, S. A. *Examiner's manual: Illinois Test of Psycholinguistic Abilities*. Champaign: University of Illinois Press, 1961.

MEYERS, C. E. What the ITPA measures: A synthesis of factor studies of the 1961 edition. *Educational and Psychological Measurement*, 1969, **29**, 867–876.

OSGOOD, C. E. A behavioristic analysis of perception and language as cognitive phenomena. In, *Contemporary approaches to cognition*. Cambridge, Mass.: Harvard University Press, 1957. (a)

OSGOOD, C. E. Motivational dynamics of language behavior. In M. R. Jones (Ed.), *Nebraska symposium on motivation*. Lincoln: University of Nebraska Press, 1957. (b)

PARASKEVOPOULOS, J. N., & KIRK, S. A. *The developmental and psychometric characteristics of the revised Illinois Test of Psycholinguistic Abilities*. Urbana: University of Illinois Press, 1969.

RYCKMAN, D. B., & WIEGERINK, R. The factors of the Illinois Test of Psycholinguistic Abilities: A comparison of 18 factor analyses. *Exceptional Children*, 1969, **36**, 107–113.

SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 1962, **27**, 125–140. (a)

SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 1962, **27**, 219–246. (b)

SIEGEL, S. *Nonparametric statistics*. New York: McGraw-Hill, 1956.

SILVERSTEIN, A. B. Variance components in the Illinois Test of Psycholinguistic Abilities. *Perceptual and Motor Skills*, 1967, **24**, 1315–1316.

UHL, N. P., & NURSS, J. R. Socio-economic level styles in solving reading-related tasks. *Reading Research Quarterly*, 1970, **5**, 452–480.

WAKEFIELD, J. A., JR., & DOUGHTIE, E. B. The geometric relationship between Holland's personality typology and the Vocational Preference Inventory. *Journal of Counseling Psychology*, 1973, **20**, 513–518.

(Received July 30, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 416–423

SOME PSYCHOMETRIC QUESTIONS IN THE EVALUATION OF PROFESSORS[1]

RICHARD M. JAEGER[2] AND TOM D. FREIJO

University of South Florida

This study investigated two research questions: (*a*) Would rewording items on a questionnaire for evaluating faculty teaching effectiveness substantially affect students' ratings? (*b*) Would students' ratings of professors' teaching quality be totally consistent with their ratings of benefits derived from courses? Subjects were 358 students in 21 classes at the University of South Florida who were administered two faculty evaluation questionnaires. Results showed that student ratings were affected very little by a major rewording of items and that a substantial degree of linear independence existed between students' perceptions of the quality of an instructional process and their perceptions of the degree to which they benefited from the instructional process.

Although scholarly publication has been a dominant criterion in judging the worth of faculty in many institutions, students (and in the case of public institutions, state legislatures) are now demanding that quality and responsibility in teaching be given dominant or, at least, substantial weight. In Florida, the Board of Regents of the state university system has developed guidelines on criteria to be used for retention, tenure, and promotion of faculty. The three areas of teaching, research, and service are to be weighted 70%, 20%, and 10%, respectively.

It is self-evident that service activities are worthwhile and readily countable. The worth of research work can be judged by the adversary system of public debate and prepublication review procedures employed by scholarly journals (although some would argue that the judgment procedure is often far too lenient). When it comes to judging teaching quality, however, we are largely at a loss. We have no general agreement on good models, and the recent review works of Rosenshine (1970) and Gage (1972) assembled little evidence on teacher behaviors that consistently predict desirable changes in students.

Scriven (1972) suggested that college professors be judged by the outcomes they produce, by their peers, and by their students. Judgment by outcomes is a radical idea in the evaluation of college professors—a procedure yet to be tried. While it presents methodological problems of enormous magnitude (e.g., standardization, equating, and the establishment of utilities), one can hope for progress in the coming years. Serious evaluation of teaching quality by a college professor's peers is effectively precluded by the closed-door ethic of the university. As a result, peer evaluation is often based upon unfounded rumor and suffers from a composite halo effect derived from the more observable activities of university faculty. So, in the world of the practical, we are left with students' judgments of a professor's competence as a teacher; such judgments are the subject of this paper.

Student judgments of a professor's teaching ability are most often secured through pencil-and-paper fixed-option questionnaires. Gustad (1961) indicated that among 584 colleges and universities surveyed, student ratings as a means of evaluating faculty were used most often. Werdell (1967) indicated that in recent years, students

[1] The authors wish to express their appreciation to Ronald Register who helped in compiling and analyzing the data.

[2] Requests for reprints should be sent to Richard M. Jaeger, College of Education, University of South Florida, Tampa, Florida 33620.

have provided a strong impetus for the use of student ratings in evaluating faculty. Eble (1970) found that student opinion questionnaires were the most commonly used form of teacher and course evaluations.

A number of student rating forms have been developed during the last five decades. Excellent brief histories of the development of these instruments, along with a description of a number of the instruments, are given by Werdell (1967) and Eble (1970). The content of these instruments is very similar; a set of questions on the behavior of the professor and the content and organization of the course is typical. The structure of teaching behaviors and course characteristics represented by questionnaries has been explored by Isaacson et al. (1964), Meredith (1969), Caffrey (1969), and McKeachie (1969), using factor-analytic techniques.

Meredith (1969) made the observation that there was need for a set of items in teacher rating scales which would measure a class of variables associated with "effects on the learner." In a more recent study, Hartley and Hogan (1972) pursued Meredith's suggestion by supplementing the traditional process-oriented questionnaire with items that addressed students' perceptions of their "progress and performance as related to this course." Upon factor analyzing the results, they found that all of the factors were defined exclusively by items from either the "process-oriented" part of the questionnaire or the "performance" part. They indicated that this clear separation of items into process and performance factors suggests the need for continued exploration and expansion of instruments for securing student evaluations of courses.

In this paper, two research questions are explored. The first is, If the items on a questionnaire for student rating of professors are reworded, will substantial changes in the rated abilities of professors result? And the second, as a partial replication of the work of Hartley and Hogan, is, If some items on a questionnaire for student rating of professors concern the process of instruction and others concern the outcomes of instruction, will the process items and the outcome items define mutually exclusive, orthogonal factors? That is, will the variables that correlate highly with some factors be exclusively of the "process" type, and will the variables that correlate highly with other factors be exclusively of the "outcome" type? Such a result would indicate a substantial degree of linear independence between the outcome variables and the process variables and, therefore, a substantial degree of exclusiveness between professors rated highly on the process variables and those rated highly on the outcome variables.

Method

Instruments

Two rating scales were used to obtain student opinions on professors, courses, and the student's own development. The first scale, an officially approved form at the University of South Florida, is composed of 17 items, 9 pertaining to the instructor of the course and 8 pertaining to the course. The items of the rating scale were as follows:

Standard Questionnaire

Please mark the following 17 items: (1) Below Average, (2) Average, (3) Above Average, (4) Superior, (5) Excellent.

Instructor of Course	1	2	3	4	5
1. Instructor's knowledge of course	—	—	—	—	—
2. Instructor's preparation for class presentations	—	—	—	—	—
3. Clarity of instructor's class presentations	—	—	—	—	—
4. Instructor's enthusiasm for course	—	—	—	—	—
5. Relationship of instructor's presentations to course objectives	—	—	—	—	—
6. Instructor's stimulation of students' thinking, consistent with course objectives	—	—	—	—	—
7. Instructor's fairness and objectivity in grading	—	—	—	—	—
8. Instructor's interest in students	—	—	—	—	—
9. Instructor's overall effectiveness	—	—	—	—	—
Course					
10. Clarity of course objectives	—	—	—	—	—

11. Relationship of course content to course objectives — — — — —
12. Organization of course — — — — —
13. Assignments appropriate to level of course — — — — —
14. Amount of assignments for level of course — — — — —
15. Evaluation appropriate to level, content, and objectives of the course — — — — —
16. Amount of evaluation for level of course — — — — —
17. Course's overall effectiveness — — — — —

The second scale was developed by the authors for this study. It too contains 17 items, 9 under the heading "the process of teaching" and 8 under the heading "the product of teaching".

Examples of items in both sections of the revised scale are displayed.

The product items on the revised questionnaire are of two types. Pairs of items deal first with the student's perception of his/her skill development as a result of taking the course and then with the student's perception of the professor's facilitation of his/her skill development. Three pairs of items are addressed to the knowledge and comprehension levels of the cognitive taxonomy developed by Bloom et al. (1956), and one pair of items concerns the acquisition of values, following the work of Krathwohl, Bloom, & Masia (1964).

In the standard questionnaire, a common 5-point scale follows each item: the first option is labeled "1, below average;" the second, "2, average;" the third, "3, above average;" the fourth, "4, superior;" and the fifth, "5, excellent." The scale is implicitly asymmetric and attempts to spread a set of ratings that have, in the past, clustered at the high end of the scale. The items are stated as short phrases but never as complete sentences. Examples include "organization of course" and "instructor's enthusiasm for course".

A. *The Process of Teaching*

24. In comparison to other instructors you have had (college) how would you rate the clarity of this instructor's presentations and lectures in class?

Among the least clear (bottom 10%) (1)	Not as clear as most (bottom 30%) (2)	About average (3)	Clearer than most (top 30%) (4)	Among the clearest (top 10%) (5)

26. Did the material presented by this instructor (both in class and through any outside assignments) correspond to the objectives stated for this course.

Not at all (1)	Very little (2)	To some degree (3)	Substantially (4)	Completely (5)

29. In comparison to other instructors you have had (college), how would you rate the *quality* of this instructor's evaluation of students' knowledge and skills?

Among the poorest (bottom 10%) (1)	Poorer than most (bottom 30%) (2)	Average (3)	Better than most (top 30%) (4)	Among the best (top 10%) (5)

32. If you had the opportunity to take another course from this instructor, would you do so?

Definitely not (1)	Probably not (2)	I don't know (3)	Probably so (4)	Definitely so (5)

B. *The Product of Teaching*

33. By taking this course, did you gain knowledge which you consider to be valuable and useful?

Definitely not (1)	Probably not (2)	Possibly so (3)	Probably so (4)	Definitely so (5)

34. To what extent was your gaining of new knowledge facilitated by the instructor?

Not at all (1)	Very little (2)	To some degree (3)	Substantially (4)	Greatly (5)

39. By taking this course, did you acquire a set of educational values which are different from the value-structure you held previously?

Definitely not (1)	Probably not (2)	Possibly so (3)	Probably so (4)	Definitely so (5)

40. To what extent was your acquisition of a new set of educational values facilitated by the instructor?

Not at all (1)	Very little (2)	To some degree (3)	Substantially (4)	Greatly (5)

All of the process items were constructed to parallel a professor rating questionnaire previously used at the University of South Florida. Five of these items have parallels in a questionnaire currently used:

Standard Questionnaire Item	Revised Questionnaire Item
Clarity of instructor's class presentations Below Average (1) ... Average (2) ... Excellent (5)	In comparison to other instructors you have had (college), how would you rate the clarity of this instructor's presentations and lectures in class? Among the least clear (bottom 10%) ... Among the clearest (top 10%)
Organization of course Below Average (1) ... Average (2) ... Excellent (5)	In comparison to other instructors you have had (college), how would you rate this instructor's organization of this course? (Did the sequence of topics and the mode of presentation help you in learning the material presented?) Among the poorest (bottom 10%) (1) ... Among the best (top 10%) (5)
Instructor's enthusiasm for course Below Average (1) ... Average (2) ... Excellent (5)	In your opinion, is this instructor's attitude toward teaching this course, as reflected by his (her) enthusiasm in classroom presentations Extremely negative (bottom 10%) (1) ... Extremely positive (top 10%) (5)
Instructor's fairness and objectivity in grading Below Average (1) ... Average (2) ... Excellent (5)	In comparison to other instructors you have had (college), how would you rate the *quality* of this instructor's evaluation of students' knowledge and skills? Among the poorest (bottom 10%) (1) ... Among the best (top 10%) (5)
Instructor's overall effectiveness Below Average (1) ... Average (2) ... Excellent (5)	In comparison to other instructors you have had (college), how would you rate the overall effectiveness of this instructor as a teacher? Among the poorest (bottom 10%) (1) ... Among the best (top 10%) (5)

The process items in the revised questionnaire, while structurally similar to those in the standard questionnaire, differ in several respects. Each item is composed of at least one complete sentence, usually in the form of a question. The symmetric 5-point rating scales make use of adjectives that are grammatically consistent and logically consistent with the item stems. The items on the standard questionnaire were revised in these ways in an attempt to make the questionnaire less ambiguous and to increase the consistency of students' interpretations of items. If this objective could be achieved, the error variance of response due to misinterpretation of questions would be reduced, and the reliability of students' responses would be correspondingly increased.

Although the adjectives associated with the scale points on the standard questionnaire are implicitly norm referenced, the population to be used in making normative judgments is not identified. In contrast, most of the process items in the revised questionnaire use the phrase, "in comparison to other instructors you have had (college)," to identify a specific reference group and to emphasize the normative basis of the judgements requested. Since the results of student ratings are often interpreted normatively, it is important that students rate their professors comparatively.

Wherever feasible, a numerical referent (bottom 10%, bottom 30%, etc.) has been associated with each scale point in the revised questionnaire, in addition to the Digits 1 through 5 associated with the adjectives in the standard questionnaire. It is likely that students interpret the numbers associated with the scale points of the standard questionnaire ordinally, since they are associated with adjectives that are ordinally arranged on a scale

of desirability. Students cannot confidently assume these numbers, and their associated adjectives, to have interval scale properties. Is "5, excellent" as far above "4, superior" as "3, above average" is above "2, average"? Such an assumption is unwarranted. Normative numerical referents were used in the revised questionnaire in an attempt to reduce the ambiguity of the scale points in the standard questionnaire.

Sample

A 50% systematic random sample of faculty in the College of Education, University of South Florida, was asked to participate in the administration of both instruments. Usable results were obtained from 21 faculty members (about a fourth of the designed sample). While one might suspect that participating faculty represent those most confident about their teaching abilities, there is no reason to suspect that lack of representativeness will differentially bias the comparison of the standard questionnaire and the new questionnaire. The responding faculty provided ratings from 358 students in 16 undergraduate courses and 5 graduate courses. The subject matter of the courses included physical education, elementary education, educational foundations, science education, and social studies education.

Procedure

Faculty participation in the study was requested by letter. Those who agreed to participate were provided with standard questionnaires and revised questionnaires for a single class. Those whose last names began with the Letters A through M were asked to administer the standard questionnaires first, to have the standard questionnaires collected, and then to administer the revised questionnaires. Those whose last names began with the Letters N through Z were asked to distribute the revised questionnaires first, to have them collected, and then to administer the standard questionnaires. Students completed both questionnaires in a single class period, under the supervision of a fellow student. Student ratings were then delivered by students to a scoring service, and faculty were informed of results several weeks after they completed student evaluations.

Results and Discussion

Our first research question, "If the items on a questionnaire for student rating of professors are reworded, will substantial changes in the rated abilities of professors result?", can be explored in several ways. If we focus on individual teaching behaviors, normative inquiries and absolute inquiries lead to different modes of analysis. The normative question can be put as follows: "If items are reworded, will there be substantial changes in the rank order of professors on rated behaviors?" Data in Table 1 bear on this question. Mean student ratings were computed for each professor on each of the paired items shown above. Means were then converted to ranks and a Spearman rank-order correlation coefficient was computed for each pair of items. The rank-order correlations range from a low of .73 for the items dealing with instructor enthusiasm to a high of .92 for the items dealing with clarity of instruction. If we square these correlations, it appears that at least half of the variation in ratings of professors' behavior on these dimensions can be predicted from reworded items. These results provide confidence in the parallelism of the paired items, show extraordinary parallel-forms reliabilities for single-item scales, and tend to convince us that the rewording of items used here doesn't make much difference in the ranking of professors. We suspect that these correlations are nearly as high as their identical-forms, test–retest reliabilities will allow.

TABLE 1

Spearman Rank-Order Correlation Coefficients for Paired Items with Indicated Descriptors

Descriptor	Correlation coefficient
Clarity of instruction	.92
Organization of course	.87
Professor's enthusiasm	.73
Quality of grading	.76
Professor's overall effectiveness	.85

When administered in the University of South Florida, College of Education, 5-point rating scales on professors' behaviors generally resulted in median ratings close to 4. This may be due to superior teaching (one might presume that education faculty practice what they preach), or it may be artifactually induced by the rating scales used. If everyone is rated above average, an average rating begins to lose meaning as an anchor point. Suspecting that the results might be explained in part by the absence of numerical referents in the scale-point anchors (below average, average, superior, ex-

cellent, etc.), we expected that the use of numerical anchors would provide a more symmetric distribution of ratings. After all, could many more than 10% of the professors be "among the best (top 10%)"? If we generalize from ratings of professors in the College of Education, this is apparently so. While our absolute results cannot be generalized to the entire college because of the self-selection of our respondents, we can compare the distribution of responses on paired items. Because the numerical scales on the two questionnaires have different points associated with average performance, we have reparameterized the distributions about their respective averages. Table 2 shows percentages of professors rated above average for these items.

While the revised questionnaire did produce slightly more symmetric distributions of ratings (a nearly 10% difference in proportions rated above average on "organization of course"), the differences were not practically significant.

In reviewing the completed questionnaires, a tendency among many students to rate a professor consistently high or consistently low on all items was noted, an evidence of a halo effect. Since it was suspected that few professors are consistently high on all behaviors, the degree to which paired items on the two questionnaires caused students to differentiate among rated behaviors was examined. A principal-components analysis of the paired items on the standard questionnaire and the paired items on the revised questionnaire was completed, and it was reasoned that a more complex factor structure would indicate a reduction in composite halo effect. Table 3 indicates the cumulative proportions of variance accounted for by successive numbers of principal components for the paired items on the standard questionnaire and the revised questionnaire. While a single factor accounts for 68% of the variance among professors on the standard items, it accounts for 60% of the variance among professors on the revised items. Again it can be seen that the difference is in favor of the revised questionnaire, but the magnitude of the difference casts doubt on its educational significance.

TABLE 2

PERCENTAGE OF 21 PROFESSORS RATED ABOVE AVERAGE ON PAIRED VARIABLES WITH INDICATED DESCRIPTORS

Descriptor	% above average	
	Standard questionnaire	Revised questionnaire
Clarity of instruction	89.5	81.3
Organization of course	91.0	81.6
Professor's enthusiasm	96.9	96.4
Quality of grading	90.9	81.4
Professor's overall effectiveness	91.9	86.1

TABLE 3

CUMULATIVE PROPORTION OF VARIANCE ACCOUNTED FOR BY INDICATED NUMBER OF PRINCIPAL COMPONENTS FOR PAIRED VARIABLES ON STANDARD AND REVISED QUESTIONNAIRES

No. principal components	Questionnaire	
	Standard	Revised
1	.68	.60
2	.80	.76
3	.89	.87
4	.95	.94
5	1.00	1.00

The second research question explored the relationships among students' ratings of professors on process variables and outcome variables. Those variables concerned with a professor's teaching behaviors and his organization and conduct of a class were termed process variables, and those variables concerned with students' perceptions of the benefits they derived from a class were termed outcome variables. Based on the work of Hartley and Hogan (1972), it was hypothesized that there would be some degree of linear independence among process variables and outcome variables. From this it follows that professors rated most highly on process variables would not necessarily be rated most highly on outcome variables.

To explore this question, a principal-components analysis of the entire revised questionnaire was completed. It was assumed that all of the variation to be fac-

TABLE 4

VARIABLE NUMBERS, FACTOR LOADINGS AND BRIEF DESCRIPTIONS OF VARIABLES LOADING HEAVILY ON FIVE FACTORS EXTRACTED FROM A MATRIX OF STUDENT RATINGS OF PROFESSORS AND COURSES

Factor	Variable number	Loading	Description of variable
1	25	.75	organization of course
	24	.74	clarity of instruction
	26	.74	relationship, materials to objectives
	30	.64	instructor's overall effectiveness
	31	.57	recommend to fellow student
2	39	.86	acquisition of new educational values
	40	.85	instructor facilitates value acquisition
3	35	.87	gain ability to relate past knowledge
	36	.77	instructor facilitates ability to relate past knowledge
4	28	.82	enthusiasm of instructor
	27	.70	willingness and availability of instructor
	31	.53	recommend to fellow student
	32	.53	take another course from this instructor
	29	.50	quality of evaluation procedures
5	38	.80	instructor facilitates comprehension of general concepts
	37	.72	learns to comprehend general concepts

Note. The variables shown are those with loadings in excess of .5. It was assumed that all variance was contained in a common factor space. Five factors accounted for 73% of the total variance.

tored was contained in a common factor space, and the varimax procedure was used to rotate the principal-components solution. A summary of results is contained in Table 4.

The first factor was defined by 5 items with loadings of at least .5. Note that all of these items concern the process of instruction—organization, clarity, sticking to objectives, overall effectiveness, and an operational judgement of overall quality. We were tempted to label this factor "judged managerial effectiveness of the professor". Factor 2 was defined by 2 related questions on students' perceptions of their acquisition of new educational values and the professor's role in facilitating that acquisition. Like Factor 2, Factor 3 was defined by students' perceptions of their progress. Only 2 items had loadings in excess of .5 on Factor 3: students' perceptions of their gain in ability to relate and integrate past knowledge, and their feelings that the professor facilitated this gain. Factor 4 can be labeled a behavioral dimension. Three of the 5 items concerned the professors' enthusiasm, willingness, availability, and fairness in grading; the other items could be characterized as operational endorsements of these qualities. Finally, the fifth rotated factor was defined by a pair of items concerned with the acquisition of general concepts and the professor's facilitation of that acquisition. The five rotated factors accounted for 73% of the variance on the 17 items.

Our findings on the relationships among process and outcome variables were consistent with those of Hartley and Hogan (1972). Process judgements and outcome judgements appear to define unique dimensions, since in no case did a process item and an outcome item load heavily on the same factor. Since the mutally orthogonal factors obtained were defined exclusively by a set of process variables or a set of outcome variables, it follows that these variables themselves exhibit a substantial degree of linear independence.

A similar interpretation of the factor-analysis results leads to the conclusion that students perceive courses as contributing to their self-development in different ways. This follows since variables concerned with specific kinds of student benefits defined mutually orthogonal factors. It is reassuring to note that students view professors as important in contributing to their course-related development.

In summary, then, our study shows that students' ratings of professors are little affected by the rewording of items used here and that the typical rating questionnaire on the process of instruction should be supplemented by a questionnaire containing outcome items.

REFERENCES

BLOOM, B. et al. *Taxonomy of Educational Objectives, Handbook I, Cognitive Domain.* New York: David McKay, 1956.

CAFFREY, B. Lack of bias in student evaluation of teachers. *Proceedings of the 77th Annual Convention of the American Psychological Association*, 1969, **4**, 641–642. (Summary)

EBLE, K. E. *The recognition and evaluation of teaching.* Salt Lake City, Utah: Project to Improve College Teaching, 1970.

GAGE, N. L. How to improve teaching. Paper presented at the meeting of the American Psychological Association, Honolulu, September 1972.

GUSTAD, J. W. Policies and practices in faculty evaluation. *Educational Record*, 1961, **42**, 194–211.

HARTLEY, E. L., & HOGAN, T. P. Some additional factors in student evaluation of courses. *American Educational Research Journal*, 1972, **9**, 241–250.

ISAACSON, R. et al. Dimensions of student evaluations of teaching. *Journal of Educational Psychology*, 1964, **55**, 344–351.

KRATHWOHL, D., BLOOM, B., & MASIA, B. *The Taxonomy of Educational Objectives, Handbook II, Affective Domain.* New York: David McKay, 1964.

MCKEACHIE, W. J. Student ratings of faculty, *American Association of University Professors Bulletin*, 1969, **55**, 439–444.

MEREDITH, G. M. Dimensions of faculty course evaluation. *Journal of Psychology*, 1969, **73**, 27–32.

ROSENSHINE, B. The stability of teacher effects upon student achievement. *Review of Educational Research*, 1970, **40**, 647–662.

SCRIVEN, M. The evaluation of university faculty. Paper delivered at the meeting of the American Association of Colleges, San Francisco, May 1972.

WERDELL, P. R. *Course and teacher evaluation.* (2nd ed.) Washington, D. C.: U. S. National Student Association, 1967.

(Received July 30, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 424-431

PREDICTION OF PRODUCTIVITY FROM PERSONALITY TEST SCORES

CHARLES F. ELTON[1] AND HARRIETT A. ROSE

University of Kentucky

This study investigated the relationship between the quantity of nonacademic achievements and personality test scores. The total nonacademic score—the dependent variable—was derived from nine nonacademic achievement scales on the American College Test. The predictor variables consisted of 14 scale scores from the Omnibus Personality Inventory (OPI) and 1 academic aptitude measure. The relationship between the predictor and dependent variables was analyzed by a stepwise regression analysis for a sample of 505 men and 496 women. In a replication, the number of students consisted of 1,329 men and 1,046 women. The Social Extroversion, Estheticism, and Impulse Expression scale scores on the OPI were the three best predictors in each analysis.

In their relatively brief history, the merit of nonacademic achievement or accomplishment scales has been examined by several investigators. The first published study using these scales was that of Holland (1961). He reported the intercorrelations between 72 predictors and 3 criterion measures in a sample of National Merit finalists. The predictors consisted of personal, demographic, and parental variables, and the criterion measures consisted of 1 measure of academic talent (high school grades) and 2 measures of nonacademic talent (scientific and artistic performance). Examples of items in these latter 2 measures are as follows: gave an original paper at a scientific meeting sponsored by a professional society and won literary award or prize for creative writing, respectively. Students were asked to check those items which applied to them. Although Holland noted the difficulties inherent in defining creativity, he labeled the scientific and artistic scales as measures of creative performance. A negligible correlation was found between academic and creativity scores. Similar correlations were reported between creativity and personality scores, the latter measured by Gough's Differential Reaction Scale and tests adapted from Barron's Complexity–Simplicity, Independence of Judgment and Originality. Median correlations were reported as .12 and .14 for boys and girls, respectively.

In an extension of this study, undergraduate achievements in scientific, artistic, leadership, and scholastic areas were predicted from data obtained in the senior year of high school for four independent samples of National Merit finalists over one-, two-, three-, and four-year college intervals (Holland & Astin, 1962). It was found that the best single predictor of each of the four kinds of college achievement was similar achievement in high school. Among the predictors were personality test scores obtained from the Sixteen Personality Factor Questionnaire, California Psychological Inventory (CPI), and an early form of the Omnibus Personality Inventory (OPI). The median correlations reported between these predictors and achievements in areas of leadership, art, and science were .12, .13, and .02, respectively.

In another related study, Holland and Nichols (1964) employed multiple regression analysis to predict college achievements including grades, leadership, science, dramatic arts, literature, music, and art, from a high school assessment of interests

[1] Requests for reprints should be sent to Charles F. Elton, Room 111, Dickey Hall, University of Kentucky, Lexington, Kentucky 40506.

goals, activities, and aptitudes for a large sample of high-aptitude students. Again, similar nonacademic accomplishments in high school were found to be significant predictors of college accomplishment.

Subsequently, the relationship between nonacademic and academic accomplishment was studied by Holland and Richards (1965), Richards, Holland, and Lutz (1967), Werts (1967), Elton and Shevel (1969), Wallach and Wing (1969), and Wing and Wallach (1971). Nonacademic talent was found to be independent of academic talent for a variety of student populations in the studies reported by Holland and his colleagues. Werts (1967), however, reported finding a relationship between nonacademic and academic achievements and suggested that the lack of an existing relationship in previous studies might have been masked by the use of the correlation coefficient as the principal analytic technique. Other investigators later supported the independence between these two types of accomplishments (Elton & Shevel, 1969; Wallach & Wing, 1969).

The relationship between personality variables and nonacademic achievements has been reported as negligible (Holland, 1961; Holland & Astin, 1962). These results, however, are subject to two possible qualifications. First, marked changes have occurred in the nonacademic achievement scales over the past ten years, that is, items have been altered in some scales and the number of scales has been expanded. Second, personality traits may predict quantity of nonacademic accomplishments, if it is assumed that nonacademic achievement scales are intercorrelated. The best evidence for this assumption is derived from an inspection of the intercorrelations between nonacademic achievements in the areas of science, leadership, art, music, speech, and writing which suggests that nonacademic talent is more general than specific in its occurrence. For example, the median correlation among nonacademic achievements for males is .37 (Holland & Richards, 1965, Table 2). The purpose of this study is to investigate the relationship between the quantity of nonacademic achievements and personality test scores.

Method

Multiple regression analysis was employed to establish the relationship between the predictor variables—personality and ability measures—and the dependent variable—total number of self-reported nonacademic accomplishments.

The predictors consisted of 14 scale scores on the OPI and a measure of academic aptitude.

OPI Tests

The OPI, Form F, contains 385 items designed to assess selected attitudes, values, and interests of college students, chiefly relevant in the areas of normal ego-functioning and intellectual activity. The specific scales of the OPI include the following measures:

1. Thinking Introversion (TI). Persons scoring high on this scale are characterized by a liking for reflective thought and academic activities.
2. Theoretical Orientation (TO). High scorers indicate a preference for dealing with theoretical concerns and problems and for using the scientific method in thinking.
3. Estheticism (Es). High scorers endorse statements indicating diverse interests in artistic matters and activities and a high level of sensitivity and response to esthetic stimulation.
4. Complexity (Co). High scorers are tolerant of ambiguities and uncertainties; they are fond of novel situations and ideas.
5. Autonomy (Au). High scorers show a tendency to be independent of authority as traditionally imposed through social institutions.
6. Religious Orientation (RO). High scorers are skeptical of conventional religious beliefs and practices and tend to reject most of them, especially those that are orthodox or fundamentalistic in nature.
7. Social Extroversion (SE). High scorers display a strong interest in being with people, and they seek social activities and gain satisfaction from them.
8. Impulse Expression (IE). High scorers have an active imagination, value sensual reactions and feelings, and are aggressive.
9. Personal Integration (PI). High scorers admit to few attitudes and behaviors that characterize socially alienated or emotionally disturbed persons.
10. Anxiety Level (AL). High scorers deny that they have feelings or symptoms of anxiety, and they do not admit to being nervous or worried.
11. Altruism (Am). High scorers are affiliative persons who are trusting and ethical in their relations with others.
12. Practical Outlook (PO). High scorers are interested in practical, applied activities

and tend to value material possessions and concrete accomplishments.

13. Masculinity–Femininity (MF). High scorers (masculine) deny interests in esthetic matters, and they admit to few adjustment problems.

14. Response Bias (RB). High scorers respond in a manner similar to a group of students who were explicitly asked to make a good impression.

The reliabilities of these scales range from .67 to .89; a complete description of their development as well as validity data may be found in the *OPI Manual* (Heist and Yonge, 1968).

Academic Aptitude

The American College Test (ACT) Composite score was used as a predictor also. This variable was included because the relationship between ability and the recently added nonacademic achievement scales of athletics, practical skills, and work experience has not been investigated.

The dependent variable was the total number of self-reported, nonacademic accomplishments taken from the Student Profile section of the ACT. These self-reported or claimed accomplishments are grouped in nine scales each with seven items. The student claims to have accomplished a given activity by responding either "yes" or "no" to that activity.

Maxey and Ormsby (1971) investigated the validity of 28 self-reported achievement items in large samples of males and females. Items were selected from the nonacademic scales which could be verified from either high school records or by high school personnel. For men, item agreement between student self-report or claim and school report ranged from a low of 70.5% for the item "actively campaigned to elect another student" to a high of 97.7% for the item "named to an all state team." These same two items defined the extreme percentage agreements among women; for example, the low was 67.7% while the high was 99.5%. These investigators interpreted their findings as indicating excellent agreement between student claim and school report for items in which schools could be expected to keep accurate records. Some items, however, reflected activities in which either the school did not keep good records or it was difficult for school personnel to ascertain whether or not the student engaged in a specific activity. It should be pointed out that Maxey and Ormsby examined the validity of only 28 items, hence the validity of 35 items is unknown.

The "yes" responses were summed over the 9 scales to provide a total, self-reported nonacademic accomplishment score. The use of a total nonacademic achievement score as the dependent variable assumed that the 63 self-reported items in the 9 scales could be treated as 1 nonacademic achievement scale. The justification for this procedure follows. First, each scale consisted of 7 items ranging from common and less important accomplishments to rarer and more important accomplishments. In the absence of any published data on the degree of intercorrelation between the 63 items, independence between the items was assumed. Second, the correlations between the 9 separate scale scores and the total number of self-reported achievements were examined. The median correlation for men was .51; for women it was .49.

Sample items from the nine self-reported nonacademic achievement scales follow:

Leadership: appointed to a student office; organized a school political group or campaign; participated in a nonschool political campaign.

Music: composed music; performed with a professional musical group (orchestra, band, choral group); performed in a school musical group.

Speech: placed first, second, or third in a regional or state speech or debate contest; entered a school speech or debate contest; had leads in high school or churchsponsored plays.

Art: finished a work of art (painting, ceramics, sculpture, etc.) on my own (not as part of a course); exhibited a work of art in a statewide or regional show; had photographs, drawings, or other artwork published in a public newspaper or magazine.

Writing: had poems, stories, essays, or articles published in school publication; wrote an original but unpublished piece of creative writing on my own (not as part of a course); won literary award or prize for creative writing.

Science: did an independent scientific experiment (not as part of a course); won a prize or award (of any kind) for scientific work or study; placed first, second, or third in a regional or state science contest.

Athletics: participated in one or more varsity athletic team events (football, basketball, baseball, etc.) while attending high school; earned a varsity letter in one or more sports in high school; received all-city, league, county, or state team award (including honorable mention).

Work experience: worked regularly for pay; earned one or more raises or promotions because of good work; did a job most people my age couldn't do.

Practical skills: paid a bill, purchased something with a money order, or balanced a checkbook; made a useful item by sewing; can repair an automobile if I have the right equipment.

The reliabilities of the first six scales range from .81 to .89 over a one-month interval and from .77 to .89 over a three-month interval (Technical Report for the ACT Assessment Program, 1973).

Freshman OPI scores for 505 men and 496

women were chosen for analysis. These students had entered as freshmen in 1971 and had persisted in college for at least five semesters. A stepwise multiple regression analysis was used, separately for men and women, to predict the total nonacademic accomplishment score.

The analyses described above were replicated with the entire entering freshman class of 1970 for whom all test data were available (*ns* = 1,329 men and 1,046 women).

In the regression analysis, the .01 level of significance was chosen for the inclusion of additional variables which reduced the residual variance.

Results

Tables 1 and 2 show the means, standard deviations, and intercorrelations between the independent and dependent variables. Results for men are above the diagonal, results for women are below. The correlations in both tables range from low to moderate. The highest correlation between an OPI scale and ability was found on the measure of Autonomy; this existed for both sexes in both tables. It is worth noting that most OPI scales correlate higher with the total nonacademic accomplishment score than does the measure of ability. The higher mean self-reported nonacademic accomplishment score found in Table 1 compared to that found in Table 2 may be a result of the different student populations, that is, persisting versus entering students.

The results of the regression analysis predicting total self-reported nonacademic accomplishments for the 1971 persisting freshmen are shown in Table 3; similar data are provided in Table 4 for the 1970 entering freshmen. In each analysis, the three best predictors of total self-reported nonacademic achievements are Social Extroversion, Estheticism, and Impulse Expression. The multiple correlations for persisting freshmen are appreciably higher than they are for entering freshmen. A question suggested by this finding is, What is the relationship between self-reported nonacademic accomplishments and persistence in college?

Discussion

Social Extroversion, Estheticism, and Impulse Expression are the three best predictors of the quantity of self-reported accomplishments for both sexes and for both entering and persisting freshmen. On the basis of these personality scales, the student with a large number of self-reported accomplishments may be described as follows: likes other people and derives stimulation from contact with others; possesses an interest in art, painting, sculpture, music, and literature; and expresses his impulses readily either through conscious thought or in overt action.

MacKinnon (1968) suggested that competencies and self-reported nonacademic achievement items such as those used as the dependent variable in this study are more promising predictors of future creative performance than are measures of intelligence. If, as MacKinnon suggests, nonacademic accomplishments are potential indicators of future creative performance, the findings in this study are supported by those reported by Parloff, Datta, Kleman, and Handlon (1968). These investigators administered the California Psychological Inventory to a sample of adolescents judged to be creative on the basis of their Science Talent Search projects. Test results on the California Psychological Inventory were obtained also for samples of creative mathematicians, research scientists, writers, and architects. Separate factor analyses of the California Psychological Inventory for the adolescent and adult samples revealed similar factor structures. Both the creative adults and the more creative adolescents, however, were distinguished by factors labeled adaptive autonomy and assertive self-assurance. Among their conclusions, these investigators noted, "Creative performance among the adolescents appears to be facilitated rather than inhibited by a measure of social skills and self-control [p. 548]." While the measures of nonacademic accomplishment used in this study cannot be equated with the quality or social importance of the measure of creativity used by Parloff et al. (1968), the correspondence between their description of the more creative adolescent and the description in this study of the high nonacademic achiever is apparent.

The congruence between the Parloff et al. (1968) findings and those in this study may

TABLE 1

MEANS, STANDARD DEVIATIONS AND INTERCORRELATIONS AMONG THE INDEPENDENT AND DEPENDENT VARIABLES FOR PERSISTING FRESHMEN

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Men (n = 505) $\bar{X}$	Men (n = 505) SD
1. Thinking Introversion		57	65	48	42	16	23	18	01	02	48	−61	−43	37	33	31	22.16	7.33
2. Theoretical Orientation	57		29	38	31	19	08	08	06	11	28	−42	01	50	39	16	19.26	5.05
3. Estheticism	62	36		48	31	12	13	30	−25	−24	28	−39	−71	07	13	36	9.84	4.76
4. Complexity	53	46	44		57	33	04	49	−30	−20	14	−67	40	−08	23	25	14.97	5.31
5. Autonomy	36	25	29	54		53	−08	32	−11	−05	14	−76	−30	−06	44	11	24.07	7.62
6. Religious Orientation	13	16	14	34	50		−23	26	−21	−15	−24	−37	−08	−25	21	03	12.51	5.03
7. Social Extroversion	20	11	12	14	−07	−10		21	43	28	54	01	−16	32	−06	35	21.86	7.55
8. Impulse Expression	28	25	36	53	36	37	21		−44	−39	−14	−16	−29	−40	−01	26	29.59	9.16
9. Personal Integration	−05	−06	−20	−24	−07	−16	39	−52		76	46	05	35	64	05	−02	31.82	10.13
10. Anxiety Level	01	08	−18	−07	04	−07	32	−37	76		30	01	42	55	10	−04	13.10	4.16
11. Altruism	33	18	15	12	02	−28	50	−26	56	44		−32	−28	51	18	20	19.15	5.53
12. Practical Outlook	−59	−39	−37	−70	−75	−38	−06	−26	−02	−08	−23		39	−12	−40	−15	16.35	5.53
13. Masculinity–Femininity	−28	13	−52	−24	−23	−02	−18	−24	22	31	−18	26		17	−03	−26	31.96	5.35
14. Response Bias	27	43	08	−11	−16	−23	29	−40	60	50	49	−02	23		20	11	13.01	4.64
15. American College Test	20	24	16	17	30	02	−03	00	−03	07	13	−26	09	12		−02	21.80	4.85
16. Total nonacademic achievement	39	28	42	31	08	05	44	35	−04	−02	20	−16	−27	12	15		17.10	9.74
Women (n = 496)																		
$\bar{X}$	23.12	16.80	12.68	14.70	24.44	11.18	24.86	25.98	32.05	12.94	23.14	14.81	23.56	12.84	21.26	15.80		
SD	7.43	5.19	4.69	5.47	7.35	4.85	7.28	9.40	10.24	4.21	5.10	5.32	4.90	4.18	4.78	8.23		

Note. Men are above the diagonal, women below the diagonal. Decimals have been omitted for Variables 1–16.

TABLE 2

MEANS, STANDARD DEVIATIONS, AND INTERCORRELATIONS AMONG THE INDEPENDENT AND DEPENDENT VARIABLES FOR ENTERING FRESHMEN

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Men (n = 1,329) $\bar{X}$	Men (n = 1,329) SD
1. Thinking Introversion		56	61	51	34	10	25	11	05	03	51	−60	−37	37	17	29	20.84	7.04
2. Theoretical Orientation	62		29	39	25	18	09	02	08	16	29	−40	08	53	28	21	18.61	5.01
3. Estheticism	60	32		48	23	11	17	29	−21	−21	28	−34	−64	10	02	34	9.74	4.73
4. Complexity	49	45	41		46	27	14	41	−19	−08	20	−62	−32	−02	09	22	15.11	4.86
5. Autonomy	38	28	27	49		48	02	22	02	05	13	−68	−17	−05	37	01	23.10	6.96
6. Religious Orientation	19	21	17	34	52		−10	29	−16	−09	−21	−28	−01	−19	12	−12	12.78	4.89
7. Social Extroversion	20	09	06	04	−09	−11		21	38	24	51	−12	−19	22	−10	59	21.13	7.02
8. Impulse Expression	20	14	31	46	26	31	13		−45	−36	−21	−03	−27	−44	−11	28	31.73	8.77
9. Personal Integration	06	06	−17	−17	00	−15	44	−50		75	45	−12	36	58	03	09	28.22	10.09
10. Anxiety Level	10	16	−14	−04	06	−06	34	−36	73		28	−14	44	50	10	05	11.74	4.42
11. Altruism	42	27	13	15	16	−17	49	−27	60	43		−38	−24	47	09	36	17.73	5.28
12. Practical Outlook	−61	−46	−33	−67	−72	−35	−04	−15	−12	−14	−38		24	−19	−27	−06	16.98	5.24
13. Masculinity-Femininity	−26	10	−52	−21	−18	−03	−14	−28	30	41	−14	21		20	09	−28	31.98	5.49
14. Response Bias	31	46	04	−08	−08	−19	30	−42	65	56	53	−12	29		16	22	11.85	4.14
15. American College Test	24	27	15	14	40	12	−09	03	02	11	17	−32	01	17		06	22.49	4.32
16. Total nonacademic achievement	31	20	28	20	07	−04	38	24	10	−10	20	−11	−13	20	09		14.00	6.55
Women (n = 1,046)																		
X	22.00	15.80	12.62	14.72	23.51	11.30	23.65	27.02	29.78	11.86	22.07	15.54	23.00	11.46	21.39	13.11		
SD	7.28	5.02	4.47	5.47	7.24	4.64	6.94	8.81	10.65	4.50	5.29	5.42	4.96	4.19	4.21	5.89		

Note. Men are above the diagonal, women below the diagonal. Decimals have been omitted for Variables 1–16.

TABLE 3

BETA WEIGHTS AND MULTIPLE CORRELATIONS FOR PREDICTING TOTAL NONACADEMIC ACCOMPLISHMENT AMONG 1971 PERSISTING FRESHMEN

Predictor	Men (n = 505)		Women (n = 496)	
	β	R	β	R
Social Extroversion	.518	.593	.375	.441
Estheticism	.223	.647	.288	.572
Impulse Expression	.131	.654	.174	.593
Theoretical Orientation	.123	.660		
Autonomy	−.106	.667		
American College Test Composite			.119	.605

be explained partially by a study reported by Skager, Schultz, and Klein (1965) on self-reported nonacademic accomplishments as measures of creativity. These investigators developed items similar in content and emphasis to those used by Holland (1961) in his original study. Rare as well as relatively common activities were included in their checklist of accomplishments. In addition, students—142 males from a state university and 150 males from a selective technological institution—were asked to describe briefly the activity checked. Quantity and quality scores for each subject were obtained. Raters examined each activity description and designated which was the most significant achievement. These were then sorted by five judges, including three psychologists, one physicist, and one artist, into one of six categories of quality. Instructions to the judges were as follows: (*a*) the achievement implies originality as opposed to replication; (*b*) it is self-initiated rather than organized or planned by someone else; (*c*) it involves official recognition and/or public performance; and (*d*) there is evidence of tangible product over learning or passive curiosity. The correlations found between quantity and quality were .44 and .29 for the state university and technological institution samples, respectively. The dependent variable in this study was one solely of quantity, but the moderate correlation between quantity and quality suggests that students who report large numbers of nonacademic accomplishments may also be reporting a few rare or unusual accomplishments.

Nonacademic accomplishments have been designated as measures of creative performance (Holland, 1961; Skager et al., 1965; Taylor & Ellison, 1972). The importance of the criterion problem in measuring creativity has been emphasized by Brogden and Sprecher (1964) who stated,

> But is the creative person the same as the productive person? Minimal productivity is probably necessary before a person's creativity can be identified and recognized by society. There may also prove to be at least a moderately high correlation between productivity and creativity [p. 156].

In view of the fact that the criterion problem has not been resolved, it would appear desirable to designate self-reported nonacademic accomplishment scales as measures of productivity rather than as measures of creativity. Thus far, the relationship between college nonacademic achievements and postcollege measures of creativity has not been investigated. It will be necessary to establish this link before nonacademic achievements in high school or college may be considered as potential predictors of creativity. Dellas and Gaier (1970) state in their review of research on creativity,

> It also appears that there is a necessity to develop creativity measures based on personality study rather than task performance. . . . That biographical items and past achievement have been rated

TABLE 4

BETA WEIGHTS AND MULTIPLE CORRELATIONS FOR PREDICTING TOTAL NONACADEMIC ACCOMPLISHMENT AMONG 1970 ENTERING FRESHMEN

Predictor	Men (n = 1,329)		Women (n = 1,046)	
	β	R	β	R
Estheticism	.239	.365	.181	.462
Social Extroversion	.282	.469	.281	.386
Impulse Expression	.154	.483	.253	.477
Response Bias	−.093	.491	.187	.506
Personal Integration	.145	.495		
Religious Orientation			−.092	.511
American College Test Composite			.072	.516

as the most efficient predictors (Taylor and Holland, 1964) does not mean investigators have no further work in this direction [p. 70].

The findings of this study indicate a relationship between self-reported task performance and personality test scores. Unanswered, however, is the relationship between measures of task performance and creativity.

REFERENCES

Brogden, H. E., & Sprecher, T. B. Criteria of creativity. In C. W. Taylor (Ed.), *Creativity: Progress and potential.* New York: McGraw-Hill, 1964.

Dellas, M., & Gaier, E. L. Identification of creativity: The individual. *Psychological Bulletin,* 1970, **73,** 55–73.

Elton, C. F., & Shevel, L. R. *Who is talented? An analysis of achievement.* (ACT Research Report 31) Iowa City, Iowa: American College Testing Program, 1969.

Heist, P., & Yonge, G. *Omnibus Personality Inventory manual.* New York: Psychological Corporation, 1968.

Holland, J. L. Creative and academic performance among talented adolescents. *Journal of Educational Psychology,* 1961, **52,** 136–147.

Holland, J. L., & Astin, A. W. The prediction of the academic, artistic, scientific and social achievement of undergraduates of superior scholastic aptitude. *Journal of Educational Psychology,* 1962, **53,** 132–143.

Holland, J. L., & Nichols, R. C. Prediction of academic and extra-curricular achievement in college. *Journal of Educational Psychology,* 1964, **55,** 55–65.

Holland, J. L., & Richards, J. M., Jr. Academic and nonacademic accomplishment: Correlated or uncorrelated? *Journal of Educational Psychology,* 1965, **56,** 165–174.

MacKinnon, D. W. Selecting students with creative potential. In P. Heist (Ed.), *The creative college student.* San Francisco, Calif.: Jossey-Bass, 1968.

Maxey, E. J., & Ormsby, V. J. *The accuracy of self-report information collected on the ACT Test Battery: High school grades and items of nonacademic achievement.* (ACT Research Report 45) Iowa City, Iowa: American College Testing Program, 1971.

Parloff, M., Datta, L., Kleman, M., & Handlon, J. Personality characteristics which differentiate creative male adolescents and adults. *Journal of Personality,* 1968, **36,** 528–552.

Richards, J. M., Jr., Holland, J. L., & Lutz, S. W. The prediction of student accomplishment in college. *Journal of Educational Psychology,* 1967, **58,** 343–355.

Skager, R. W., Schultz, C. B., & Klein, S. P. Quality and quantity of accomplishments as measures of creativity. *Journal of Educational Psychology,* 1965, **56,** 31–39.

Taylor, C. W., & Ellison, R. L. Predictors and criteria of creativity: A Utah progress report. In C. W. Taylor (Ed.), *Climate for creativity.* New York: Pergamon Press, 1972.

Technical Report for the ACT Assessment Program. Iowa City, Iowa: American College Testing Program, 1973.

Wallach, M. A., & Wing, C. W. *The talented student.* New York: Holt, Rinehart & Winston, 1969.

Werts, C. E. The many faces of intelligence. *Journal of Educational Psychology,* 1967, **58,** 198–204.

Wing, C. W., & Wallach, M. A. *College admissions and the psychology of talent.* New York: Holt, Rinehart & Winston, 1971.

(Received August 6, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 432-438

CLASSIFICATORY BEHAVIORS OF LOW-SOCIOECONOMIC-STATUS CHILDREN[1]

GORDON K. NELSON
Pennsylvania State University

HERBERT J. KLAUSMEIER[2]
Wisconsin Research and Development Center for Cognitive Learning University of Wisconsin—Madison

Ninety-six low-socioeconomic-status students in Grades 5, 8, and 11 were presented two classification tasks related to mathematics. Task 1 required students to state how geometric forms presented successively were alike and different. The bases on which students indicated likenesses and differences were classified as perceptible, attribute, nominal, or fiat. Task 2 was a free-sorting exercise in which the students arranged 26 geometric forms successively into seven groups. Unlike middle- and high-socioeconomic-status students, the low-socioeconomic-status students classified significantly more on the basis of perceptible likenesses and differences among concept examples than on the more mature bases of the defining attributes and the names of concepts. Low socioeconomic status and lack of instruction were related to the patterns of conceptual development.

Bruner, Goodnow, and Austin (1956) and later researchers have shown that individuals learn about and deal with their complex world by rendering things equivalent; that is, the individual learns to treat many things as equivalent by identifying certain common properties while ignoring other distinguishable characteristics. The treating of things as equivalent, usually referred to as classificatory behavior or equivalence formation, cannot be undervalued for its practical significance in educational and other settings. Understanding the bases by which maturing individuals classify things is of significance to the science of human learning and also to the design of classroom instructional practices.

According to Klausmeier, Ghatala, and Frayer (1974), the lowest level of classification involves generalizing that at least two different objects, events, or processes are alike in some way and discriminating that these same two things are different from other things. The highest, exhaustive level of classification enables the individual to identify the things he encounters as examples and nonexamples of any particular concept he has formed. Children learn to classify from the lowest to the highest levels from early childhood through adolescence.

In a somewhat analogous manner and from a developmental point of view, Piagetian theory holds that "classification implies a relation of resemblance between members of the same class, and one of dissimilarity between members of different classes [Inhelder & Piaget, 1964, p. 5]." Classificatory operations arise through structural development and represent an important aspect in the attainment of concrete operational thought (Flavell, 1963).

More explicitly than Piaget, other psychologists have emphasized the importance of experiential and cultural influences in learning to classify (Bruner et al., 1966; Evans & Segall, 1969; Klausmeier et al.,

[1] This study was supported by the Wisconsin Research and Development Center for Cognitive Learning pursuant to Center No. C-03/Contract OE 5-10-154 with the U. S. Office of Education, Department of Health, Education, and Welfare. No official endorsement by the Office of Education should be inferred.

[2] Requests for reprints should be sent to Herbert J. Klausmeier, Wisconsin Research and Development Center for Cognitive Learning, University of Wisconsin, 1025 West Johnson Street, Madison, Wisconsin 53706.

1974; Schmidt & Nzimande, 1970; Suchman, 1966). In particular, Maccoby and Modiano (1966) and Greenfield, Reich, and Olver (1966) have reported substantial differences in the classificatory behaviors of groups drawn from various cultures but have noted also that differences among these groups tend to decrease with increasing educational experience and urban influence.

The present experiment was designed to learn more about the classificatory behaviors of low-socioeconomic-status children from three age groups. It was carried out as the last of two experiments to ascertain whether the changing bases of classificatory behavior in low-socioeconomic-status children were comparable to those reported in middle- and high-socioeconomic-status children by Wiviott (1970).

Wiviott (1970) patterned the first experiment after studies by Olver (1961) and Rigney (1962). Wiviott devised two tasks for the experiment. The first task consisted of a successive presentation of instances of eight geometric concepts–square, rectangle, rhombus, parallelogram, quadrilateral, triangle, circle, and cube. The instances were presented either figurally with line drawings or verbally with the names of the eight concepts. One half of the subjects were presented eight cards, each with a drawing of a concept instance printed on the card, while the other half were presented eight cards, each with the typewritten name of the concept. In the experiment, subjects were stratified according to three grade levels—5, 8, and 11—and two levels of mathematics achievement—above and below the median score of the grade group—so that one half the subjects high in mathematics achievement at each grade level received each treatment as also did one half the subjects low in achievement.

In accordance with Olver's (1961) methods, the subject was shown the first two items, *square* and *rectangle*, and was asked how they were alike. Next, *rhombus* was presented and the subject was then asked how this differed from the first two items and how all three were alike. This continued until all items were administered, with *cube* representing the final contrast item. On this item, the subject was asked only how the figure was different from the earlier ones.

Rigney's (1962) methods served as a prototype for Wiviott's (1970) second task. The same concepts as in the first task were used with the exception of cube. A 26-card array of drawings was constructed, using instances of the concepts that varied in size and orientation. The subject was instructed to form successive groups of drawings that were alike in some way and was then asked how they were alike. After the subject completed a grouping and explained the rationale underlying it, the cards were replaced in their original array and the subject was asked to form another different group. This procedure continued until the subject had formed seven groups.

Responses of the subjects on Tasks 1 and 2 were recorded during the experiment. Subsequently each response was categorized as perceptible, attribute, nominal, or fiat.

Wiviott (1970) reported that the method of presenting the instances, the grade level of the subjects, and the achievement level were all related significantly to the bases of classification employed by the subjects. Subjects who were presented the examples figurally gave more perceptible responses than those who were given the examples verbally. An increase in grade level from fifth to eighth to eleventh was accompanied with a decrease in the perceptible basis of classification and an increase in the attribute and nominal bases. This result agrees with the conclusions of Olver (1961) and Rigney (1962). Wiviott also found that higher achievers in mathematics made fewer perceptible responses than lower achievers. Thus, Wiviott supported the earlier conclusions of Olver and Rigney by adding very important information about grade level and achievement level, which are of importance in educational settings.

The present study was undertaken to ascertain whether the bases of classification found by Wiviott in middle- and high-socioeconomic-status children also held for low-socioeconomic-status children. The grade level of the subjects, the experimental ma-

terials, and the experimental procedures were kept identical to Wiviott's experiment so that the performance of the two socioeconomic groups could be related. The present experiment is now explained more fully.

Method

Subjects

Ninety-six low-socioeconomic-status students from various schools in a city of 36,000 population served as subjects in this study. They were enrolled in the fifth, eighth, and eleventh grades—32 subjects in each grade. The total population of low-socioeconomic-status subjects enrolled in these grades was initially identified by the school principals in consultation with the school counselors. Between 60 and 80 children at each grade level were identified. Next, the occupation of the head of the household in the homes of each of the children was rated by the first author of this paper using the Warner (1960) 7-point occupation scale. Students assigned occupational ratings in Categories 6 and 7 were judged to be suitable for the study; the few who received ratings of 5 were not included. The 32 subjects at each grade level who participated in this study were selected randomly from this low-socioeconomic-status population of students.

It might be well to note that Wiviott conducted her experiment in a different city of the same state. Eighty-five percent of the parents of her subjects were in occupational Categories 1, 2, and 3, and 15% were in Categories 4 and 5. Thus there was no overlapping in the occupational categories of the two groups of subjects.

In the present experiment, the subjects were randomly assigned within each grade level to one of two methods of presentation—figural or verbal—with an equal number of male and female subjects being assigned to each method. Thus, there were 16 subjects in each possible combination of grade level (5, 8, and 11) and method of presentation (figural and verbal).

Experimental Materials

The materials used in this experiment were the same as reported earlier in the study of Wiviott (1970). The cards used in Task 1 represented instances of eight geometric concepts—square, rectangle, rhombus, parallelogram, quadrilateral, triangle, circle, and cube. The instances consisted of either line drawings or printed names of the concepts. The 26 cards used in Task 2 consisted of drawings, using instances of the eight concepts that varied in size and orientation.

Procedure

Tasks 1 and 2 were administered consecutively and individually to subjects. About 20 to 30 minutes were required for each administration. Testing was conducted in private rooms in each of the schools. As in Wiviott's experiment, responses of subjects in Task 1 were recorded on audiotape while in Task 2, a verbatim written record was kept of the cards selected for each sort and of the responses subjects gave regarding the basis for their sorting.

Procedures for the first task were patterned after the studies of Olver (1961) and Wiviott (1970). The first two cards—square and rectangle—were placed on a table before the subject, and the subject was asked to explain how the two were alike. The third card, rhombus, was then presented, and the subject was asked to explain how it was different from the first two and how they were all alike. This procedure continued until all cards were administered, with cube functioning solely as a contrast item. In all, there were six questions involving likenesses and six questions involving differences among the concept instances. For the verbal method of presentation, the names of concepts were routinely pronounced for the subjects. For the figural method of presentation, the names were withheld.

Immediately following Task 1, the 26-instance array of drawings for Task 2 was given. This free-sort exercise was modeled after the methods of Rigney (1962) and Wiviott (1970). The cards were laid out on a table before the subject. Each subject was directed to examine the cards closely, to form a group which seemed alike in some way, and to tell the basis for grouping. After the response was recorded, the subject was requested to form another group, continuing the procedure until seven groups were formed.

In addition to the preceding tasks, each subject was also given a test to determine his knowledge of the geometric forms used in the tasks. Also, information was secured concerning whether the subject had received instruction in geometry.

Treatment of the Data

All responses in Task 1 were categorized using Wiviott's (1970) four bases of classification: perceptible, attribute, nominal, and subject fiat. Responses concerning similarities were tabulated separately from responses concerning differences. In Task 2, the fiat category was not used, since Wiviott noted that it rarely occurred in the free-sort exercise. A system for categorizing Task 1 and 2 responses can be seen as follows:

1. Perceptible: The child may render the items equivalent on the basis of immediate phenomenal qualities, such as color, size, shape, or on the basis of position in time or space.

Example: They are alike because they are both black figures on white cards.
They are both printed in black ink.
The lines are straight, not slanted.
They are tilted to the right.
This one is round.
One is longer than the other.
They are diamond-shaped.

2. Attribute: The child renders the items equivalent or diverse by naming a specific attribute of the concept.

Example: They all have four sides.
They are closed figures.
They are plane figures.
They are made of line segments.

3. Nominal: The child may group items by giving a name that exists ready-made in the language. A supraordinate concept name is used as the basis of grouping.

Example: They are all parallelograms.
They are diamonds.
Both the square and the rectangle are rectangles.
They are all geometric figures.

4. Subject Fiat: The child may merely state that the items are alike or are the same without giving any further information as to the basis of his grouping, even when he is prodded.

Example: They are alike.
They are just different.

The data were analyzed in the framework of a 2 × 3 multivariate design with method of presentation (figural or verbal) included as the independent variable and grade level (5, 8, or 11) included as a stratifying variable. The total number of responses in each classification category for Tasks 1 and 2 represented the dependent variables for this experiment. Three multivariate analyses of variance were performed for each task and for the two tasks combined: one for testing differences in types of classificatory responses elicited (perceptible, attribute, nominal), another for testing effects due to grade level (5, 8, and 11), and a third for testing effects due to method of presentation (figural and verbal). The fiat classification category of Task 1 was analysed separately from the main data.

Mean scores were computed for the subjects who indicated they had received instruction in geometry and for those who indicated they had received no instruction. Mean scores were also computed on the geometry test for the low-socioeconomic-status children in the present experiment and the high-socioeconomic-status children of Wiviott's experiment.

Results

An interrater reliability check was performed on a random sample of 24 protocols, 4 subjects from each cell. The protocols were scored by 2 independent raters using the scoring format as outlined. The percentage of agreement between the two ratings was 88.2% for response data on Task 1 and 89.9% for response data on Task 2.

Table 1 gives the mean number of responses in each classification category as a function of grade level and method of presentation for Tasks 1 and 2 separately and for Tasks 1 and 2 combined. Statistically significant differences were obtained among the categories of responses when collapsing across all subgroups. In Task 1, the mean number of perceptible responses (4.90) was significantly higher than attribute responses (4.57), which in turn was higher than nominal responses (.87) ($F = 150.82$, $df = 2/89$, $p < .0001$). For Task 2, the mean number of perceptible responses (3.30) was significantly higher than attribute responses (1.39), while nominal responses (2.32) were higher than attribute responses ($F = 22.17$, $df = 2/89$, $p < .0001$). The mean number of perceptible responses (8.19) for Tasks 1 and 2 combined was significantly higher than attribute responses (5.96), and the mean number of attribute responses was higher than nominal responses (2.16) ($F = 63.34$, $df = 2/89$, $p < .0001$).

Grade level was found to be not significantly related to the bases of classification employed by subjects, either in Task 1 (F

TABLE 1

Mean Number of Responses in Each Classification Category by Subgroups for Tasks 1 and 2 and for Tasks 1 and 2 Combined

Subgroup	Task 1				Task 2			Tasks 1 and 2 (combined)		
	Perceptible	Attribute	Nominal	Fiat	Perceptible	Attribute	Nominal	Perceptible	Attribute	Nominal
Grade 5	5.53	4.31	.84	1.31	3.78	1.34	1.88	9.31	5.63	2.72
Grade 8	4.42	5.47	.56	1.59	2.84	1.69	2.47	7.06	7.16	3.03
Grade 11	4.78	3.94	1.19	2.13	3.25	1.13	2.63	8.03	5.06	3.81
Pictorial	4.71	4.60	.81	1.79	3.46	1.42	2.13	8.17	6.00	2.94
Verbal	4.98	4.54	.92	1.56	3.13	1.35	2.52	8.10	5.90	3.44
All subjects	4.90	4.57	.87	1.68	3.30	1.39	2.32	8.19	5.96	2.16

= 1.81, $df = 6/176$, $p < .09$) or in Task 2 ($F = 1.68$, $df = 6/176$, $p < .15$) when the tasks were treated independently. However, grade level was found to be significant when data from both tasks were combined. Univariate F statistics showed that the perceptible and attribute bases of classification were significantly different with respect to grade level for the perceptible category ($F = 3.84$, $df = 2/90$, $p < .03$) and for the attribute category ($F = 3.12$, $df = 2/90$, $p < .05$). Fifth graders classified relatively more than eighth graders on perceptible bases (fifth = 9.31, eighth = 7.06), whereas eighth graders classified relatively more than fifth graders on attribute bases (fifth = 5.63, eighth = 7.16). Surprisingly, however, fewer attribute and more perceptible responses were given by the eleventh graders as compared to the eighth graders (eleventh: perceptible = 8.03, attribute = 5.06). This anomalous finding is examined later in the Discussion section of this paper. The mean number of nominal responses increased with grade level although not significantly (fifth = 2.72, eighth = 3.03, eleventh = 3.81).

The subject-fiat category of classificatory responses in Task 1 was analyzed separately. This category was used to designate subjects' responses which did not fit into one of the other three categories. (Subjects would customarily respond with the words "I don't know.") The mean number of subject-fiat responses was not significantly different for the three grade levels or for the two methods of presentation.

Method of presentation was found to be not significant in Task 1, apparently because of the preponderance of perceptible responses. Therefore, different from Wiviott's results, the classifying behaviors of subjects given figural stimuli and verbal stimuli were relatively the same.

Last, no significant interactions in the bases of classifying occurred between grade level and method of presentation.

The most interesting findings from this study are that low-socioeconomic-status children classified primarily on perceptible bases and that eleventh graders used the perceptible basis as much as the eighth graders. Information concerning whether the children had received instruction dealing with geometric forms was considered in light of these findings.

At the end of the testing sessions, each subject was asked whether he had received any prior classroom instruction dealing with the geometric concepts employed in this study. At the fifth-grade level, 25% of the subjects reported having prior instruction; at the eighth-grade level, 41% said they had; but at the eleventh-grade level, no subject reported having any prior instruction. School records confirmed what was reported in the questionnaires by subjects. No eleventh-grade subject in this study had previous instruction in geometry according to school records. Curriculum consultants for the school district also stated that the fifth- and eighth-grade subjects, depending on the school attended, were affected by an upgrading in the elementary school mathematics program which in recent years has included geometric concepts. Eleventh-grade subjects were unaffected by this curriculum change. In contrast to this lack of instruction, all eleventh-grade subjects participating in the study by Wiviott had prior geometry instruction.

The categorizing responses of the subjects who had and had not received instruction were analyzed. Subjects who indicated previous instruction performed differently on both tasks than subjects who did not have this experience. Perceptible responses were fewer, while attribute responses were greater for subjects with a geometry background (geometry: perceptible = 5.62, attribute = 9.15; no geometry: perceptible = 6.11, attribute = 7.89).

Subjects in both experiments were administered the same test of geometric concepts. This test was constructed according to a paradigm outlined by Frayer, Fredrick, and Klausmeier (1969) and consisted of 56 items designed to test mastery of the concepts. Results of the concept-mastery test showed much higher mean scores for the high-socioeconomic-status children as compared to the low-socioeconomic-status children at each of the three grade levels (high

socioeconomic status: fifth = 33.3, eighth = 37.9, eleventh = 47.2; low socioeconomic status: fifth = 21.4, eighth = 28.8, eleventh = 28.4). These mean scores are consistent with the patterns of classificatory behaviors reported on the two tasks and suggest that the various bases of classifying are directly related to the level of concept mastery; that is, mastery increases as the attribute and nominal bases are used and remains low and constant when the perceptible basis of classifying is employed.

Discussion

The present experiment was undertaken as the second in a series to ascertain whether the classificatory behaviors of low-socioeconomic-status children were the same as those of middle- and high-socioeconomic-status children. One important finding of the two studies concerns the more immature bases of classification used by the low-socioeconomic-status children who showed a greater mean number of perceptible responses than the high-socioeconomic-status children at all three grade levels (low socioeconomic status: fifth = 9.31, eighth = 7.06, eleventh = 8.03; high socioeconomic status: fifth = 5.56, eighth = 4.00, eleventh = 2.59). The low-socioeconomic-status children also employed fewer attribute responses than the high-socioeconomic-status children (low socioeconomic status: fifth = 5.63, eighth = 7.16, eleventh = 5.06; high socioeconomic status: fifth = 7.38, eighth = 8.25, eleventh = 9.00) and gave fewer nominal responses (low socioeconomic status: fifth = 2.72, eighth = 3.03, eleventh = 3.81; high socioeconomic status: fifth = 4.19, eighth = 5.13, eleventh = 7.09).

Another unexpected outcome of the present study was the absence of developmental change. The expected decrease in the use of the perceptible basis of classification and the related increase in the use of the attribute basis of classification did occur from Grades 5 to 8. The eleventh-grade group, however, did not follow this pattern.

Relating these experimental findings to instruction and to lack of instruction indicated that the low-socioeconomic-status children who had received instruction used the perceptible basis less and the attribute and nominal bases more than those who had not received instruction. Also, it was shown that there was no overlap in the mean scores of the low- and the high-socioeconomic-status children on a test of geometric concepts. Here the fifth-grade high-socioeconomic-status group had a higher mean score (33.3) than the eighth-grade low-socioeconomic-status (28.8) group, which was the highest mean score of the three low-socioeconomic-status grade groups.

It appears that a combination of low socioeconomic status and lack of instruction is associated with immature conceptualizing behaviors and low achievement. Apparently low-socioeconomic-status children who do not receive instruction remain at an immature level of conceptualizing, using the perceptible properties of the concept examples rather than their defining attributes and names. Concept attainment for these children remains at a relatively low level.

These conclusions are of substantial theoretical interest and practical significance. On the theoretical side, it is noted that the orderly change in the bases of classifying found by Olver (1961), Rigney (1962), and Wiviott (1970) did not occur in the low-socioeconomic-status children, particularly between Grades 8 and 11. As Bruner et al. (1966) conjectured, it appears that instruction is in fact related not only to the particular concepts that individuals attain but also to their means of conceptualizing. Other investigators have also reported considerable differences between socioeconomic groups in classificatory behaviors (Findlay & McGuire, 1957; Raven, 1967; Siller, 1957). Some have also noted that educational experience tends to mitigate these differences (Evans & Segall, 1969; Schmidt & Nzimande, 1970).

On the practical side, it is probable that the eleventh-grade children of low socioeconomic status in this study will never function at a reasonably mature level with respect to being able to classify geometric forms and to use concepts involving geometric forms. Their classificatory behaviors undoubtedly will continue to be typical of

younger elementary school age children. This appears unfortunate in view of the fact that the younger low-socioeconomic-status subjects who received instruction were beginning to operate at the more mature levels.

REFERENCES

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. *A study of thinking*. New York: Wiley, 1956.

BRUNER, J. S., ET AL. *Studies in cognitive growth*. New York: Wiley, 1966.

EVANS, J. L., & SEGALL, M. H. Learning to classify by color and by function: A study of concept-discovery by Ganda children. *Journal of Social Psychology*, 1969, **77**, 35–53.

FINDLAY, D. C., & MCGUIRE, C. Social status and abstract behavior. *Journal of Abnormal and Social Psychology*, 1957, **54**, 135–137.

FLAVELL, J. H. *The developmental psychology of Jean Piaget*. Princeton: Van Nostrand, 1963.

FRAYER, D. A., FREDRICK, W. C., & KLAUSMEIER, H. J. A schema for testing the level of concept mastery. (Working Paper 16) Madison: University of Wisconsin, Research and Development Center for Cognitive Learning, 1969.

GREENFIELD, P. M., REICH, L. C., & OLVER, R. R. On culture and equivalence: II. In J. S. Bruner et al., *Studies in cognitive growth*. New York: Wiley, 1966.

INHELDER, B., & PIAGET, J. *The early growth of logic in the child*. New York: Harper & Row, 1964.

KLAUSMEIER, H. J. Cognitive operations in concept learning. *Educational Psychologist*, November 1971.

KLAUSMEIER, H. J., GHATALA, E. S., & FRAYER, D. A. *Conceptual learning and development: A cognitive view*. New York: Academic Press, 1974.

MACCOBY, M., & MODIANO, N. On culture and equivalence: I. In J. S. Bruner et al., *Studies in cognitive growth*. New York: Wiley, 1966.

OLVER, R. R. A developmental study of cognitive equivalence. Unpublished doctoral dissertation, Radcliffe College, 1961.

RAVEN, R. J. The development of classification abilities in culturally disadvantaged children. *Journal of Research in Science Teaching*, 1967, **5**, 224–229.

RIGNEY, J. C. A developmental study of cognitive equivalence transformations and their use in the acquisition and processing of information. Unpublished honors thesis, Radcliffe College, 1962.

SCHIMDT, W. H. O., & NZIMANDE, A. Cultural differences in color/form preference and in classificatory behavior. *Human Development*, 1970, **13**, 140–148.

SILLER, J. Socioeconomic status and conceptual thinking. *Journal of Abnormal and Social Psychology*, 1957, **55**, 365–371.

SUCHMAN, R. G. Cultural differences in children's color and form preference. *Journal of Social Psychology*, 1966, **70**, 3–10.

WARNER, W. L. Social class in America. *A manual of procedure for the measurement of social status*. New York: Harper & Row, 1960.

WIVIOTT, S. P. Bases of classification of geometric concepts used by children of varying characteristics. (Tech. Rep. 143) Madison: University of Wisconsin, Research and Development Center for Cognitive Learning, 1970.

(Received August 8, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 439–447

SENSITIVITY AND VALIDITY OF LEARNING-POTENTIAL MEASUREMENT IN THREE LEVELS OF ABILITY[1]

ELISHA Y. BABAD[2] AND MILTON BUDOFF[3]

Research Institute for Educational Problems, Cambridge, Massachusetts

Learning-potential assessment is hypothesized to be more sensitive than traditional IQ tests in tapping the intellectual potential of disadvantaged children. The Series Learning Potential Test was administered three times to bright–normal, dull-to-average and subnormal children with training in problem-relevant strategies interpolated following the second administration. Both low-IQ groups gained more than the high-IQ group from the training, and the dull group gained more than the other groups from repeated administrations without training. IQ was found to predict teacher ratings of school achievement for all groups and the Series Learning Potential Test was found to predict teacher ratings for the bright group; however, the Series Learning Potential Test was superior to IQ as a predictor in the dull-to-average and the subnormal groups. Substantial proportions of subnormal subjects reached the average reasoning level of their nonretarded peers following the short training session.

Children from poor and/or nonwhite homes tend to score at below-average levels on tests which purport to measure intelligence. Jensen (1969) has argued that the mean difference of 15 IQ points between white and black groups represents a real difference in inborn general ability. However, the IQ difference has also been frequently explained by the handicaps that poor and/or nonwhite children bring to the testing situation. They are fearful of the testing process, expect to do poorly, are often insensitive to speed requirements, are unfamiliar with the problem contents, and do not develop spontaneously the most effective strategies (by middle-class criteria) to solve the problems.

IQ tests measure the degree to which children have spontaneously acquired from their natural environment the skills and knowledge which cumulatively predict academic school success. The plausible assumption is made that a child who learned informally prior to entering school will continue to learn—formally and informally—in and out of school. Children from non-middle-class homes who do not have an equal and frequent access to school-preparatory experiences tend to score poorly on IQ tests and are often viewed as "less intelligent."

Yet, many of these same low-IQ children are competent problem-solvers in their nonschool environment, having mastered the skills, knowledge, and strategies necessary to maintain a successful adjustment. These children do learn and profit from relevant experiences more successfully than their IQ scores and school achievements indicate. In Hunt's (1961) terms, this difference in competence may represent the problem of the match between the environmental demands of the school and its tests, and the child's existing schemata in his familiar world. Assessment procedures must be developed

[1] This research was supported by Grant OEG-0-8-080506-4597 from the Research Branch of the Bureau of the Handicapped, U.S. Office of Education, and Grant MH RO1 18553 from the National Institute of Mental Health. The authors wish to extend their warm gratitude to Kathleen T. Wessling, whose assistance helped carry out this study, and to Bina Presser, who helped put everything together.

[2] Is now at the School of Education, Hebrew University of Jerusalem, Israel.

[3] Requests for reprints should be sent to Milton Budoff, Research Institute for Educational Problems, 12 Maple Avenue, Cambridge, Massachusetts 02139.

which will optimize the match and result in more culture-fair measurement of general ability.

Eells, Davis, Havighurst, Herrick, and Tyler (1951) developed problem games that were deemed relevant to the experiences of low-socioeconomic-status children, but found no change in the gap between the different social class groups. This may have been due to the demand that the child go about solving the problems as middle-class psychologists thought he should. Changing the contents of the problems was not sufficient to enable the lower-class children to narrow the gap between their problem-solving styles and those expected by the middle-class bias of the test constructors.

Budoff (1969) has described a learning-potential procedure for assessing ability among low-IQ, low-socioeconomic-status children based on a process-oriented conceptualization of intelligence. The focus of the learning-potential paradigm is on the child's trainability, that is, his ability to improve performance on reasoning problems following a systematic learning experience. Reasoning is viewed as the critical ability. The reasoning tasks are administered in a "test–train–test" sequence. Training allows the low-socioeconomic-status child to understand how to solve the problems when the contents of the problems may be strange and the appropriate strategies are not readily apparent to him.

Training is hypothesized to be particularly critical for the child from a poor and/or nonwhite background, who may learn different cognitive strategies in different expressive formats than those presumed to be available by the tests. The training helps the child to narrow the cognitive gap between his previously learned problem-solving strategies and those implicit to the problems he must ordinarily solve on the middle-class-bias tests he encounters. The test–train–test paradigm also minimizes the artificiality of the test situation. The repeated contacts with the materials in a context of support and teaching allow the school-failing child to develop a sense that he can be competent. Without this competence boost, he tends not to perform at his best, implicitly expecting failure (Zigler, 1966). The essence of this assessment strategy, then, is to impose some control on the potentially negative effects on his test performance of prior life experiences.

The pretest reflects the subject's present level of functioning and his existing ability to work with the problems. Posttraining scores reflect the child's ability under optimized conditions in which all subjects are familiar with the task and its demands, have had success in solving problems similar to those on the test, and have had the opportunity to learn and apply relevant strategies. Budoff and Corman (1974) have shown that educable mentally retarded's (EMR) pretest scores correlate highly with Stanford-Binet and Wechsler scores and other indices of psychosocial vulnerability associated with socioeconomic status. Posttraining scores were correlated only with performance scale scores.

The problems employed in the learning-potential tests fall within Jensen's (1969) Level II category (conceptual learning). According to Jensen, there is little overlap between the Level II curves of lower- and higher- socioeconomic-status groups. Jensen's theory is based on the spontaneous responses evoked in traditional intelligence testing. Since these testing procedures underestimate the potential Level II ability of substantial proportions of disadvantaged children, one would expect to find greater overlap between the curves of lower- and higher- socioeconomic-status groups on learning-potential posttests than on learning-potential pretests and other tests which tap the students' spontaneous productions. Complete overlap between the Level II curves of middle- and low-socioeconomic-status children is not assumed. Inclusion of training procedures within ability testing may identify the numerous false positives among the disadvantaged (Ortar, 1960).

To date, nonverbal reasoning tasks have been used since these are not dependent on language, which represents one major area of these children's difficulties. Three nonverbal tasks have been employed in assessment of learning potential: an altered version of the Kohs Block Designs (Budoff,

1969), Raven's Progressive Matrices (Budoff, Corman, & Litzinger, 1973), and the Series Learning Potential Test (Babad & Budoff, 1971). The Series Learning Potential Test, used in the present study, is a group test for the primary and elementary grades. Its major task is the completion of series of pictures or geometric forms, arranged in a pattern in which the figures change systematically. Each item presents a horizontal row of cells each of which contains a stimulus figure. One cell is left blank. The subjects must identify among the multiple choices on the right, the picture which best completes the series. Four concepts may vary in a series: semantic content (meaningful or geometric figures), size (large/small), color (black/white), or orientation (up/down, or left/right). The concepts may vary symmetrically or asymmetrically and the blank space may be placed in any part of the series. The hardest items are those in which the blank space appears early in the series, and the concepts change asymmetrically.

Two, equivalent, 65-item forms of the test are used, with items corresponding in concepts but with different pictures in similar arrangements. None of the 17 coaching items is identical to any of the 130 test items. The test contains several types of items. Forty items consist of meaningful pictures ranging from easy items (2 concepts are arranged symmetrically and the blank space appears at the end of the series) to more difficult ones (asymmetric alternation of 3 concepts, blank space in initial position). Improvement in solving the meaningful picture series on the posttest indicates the child can solve nontrained instances of the trained items. Ten items with geometric figures discern whether the subject can transfer the strategies learned in training with meaningful pictures to dissimilar stimuli. Fifteen double-classification items (5 matrices in meaningful pictures, 10 in geometric symbols) presented in a matrix format tests for generalization of the learned strategies to problems that require the same reasoning process in a different arrangement. The strategies may also be mediated differently.

The general hypothesis of the two studies reported is that training increases the sensitivity of the general ability test with low-IQ, low-achieving students. Thus, bright, middle-class subjects of the same chronological age should not demonstrate marked gains following practice or training since they perform at optimal levels on the pretest. They either know the solutions or know how to develop the solutions to the reasoning problems prior to training. These children's pretraining scores should therefore approximate their posttraining scores, and the dispersion of scores should decrease since training should help the laggards perform better on the task. Poor, low-IQ students, who suffer from the middle-class bias of the IQ test, should do poorly on the trial prior to training. Training should increase their mean scores and the dispersion, identifying those who are potentially able but who have failed because of the differences between their native and the school environments.

Second, despite the large differences in IQ between the groups, it was hypothesized that some proportion of the low-IQ children would attain the pretest level of their middle-class peers following training. Training is hypothesized to facilitate "induced acquisition" of the problem-relevant strategy, which compensates for the middle-class children's spontaneous acquisition. Support for this hypothesis would also indicate that some low-IQ children are not inferior to their middle-class peers in potential ability and that some IQ-defined EMRs are *educationally* rather than *mentally* retarded.

Both studies include students from three IQ groups: (*a*) bright, normal middle-class children (IQs above 100); (*b*) blue-collar, dull-to-average children (IQ, 80 to 99); and (*c*) poor, subnormal, educable mentally retarded children (IQs below 80). All students were given the Series Learning Potential Test three times in a "test–test–coach–test" sequence. In this design, the practice effect ($T_2 - T_1$), the training effect ($T_3 - T_2$), and the combined effect ($T_3 - T_1$) can be separated and examined independently.

Five specific hypotheses were tested:

1. Low-IQ children gain from repeated, uncoached exposures to the Series Learning Potential Test more than high-IQ children (practice effect, to be tested in Study 1).

2. Low-IQ children gain from training on problem-relevant strategies more than high-IQ children (training effect, to be tested in Study 1).

3. Posttraining scores predict school achievement of low-IQ groups better than group IQ scores do. This is not true for the high-IQ group (to be tested in Study 2).

4. Posttraining scores of the low-IQ group predict their school achievement better than pretraining scores. For the high-IQ group, however, pretraining scores predict school achievement as well as posttraining scores.

5. Following training, substantial proportions of IQ-defined EMR subjects will attain the mean pretest score of their nonretarded peers (to be tested in both studies).

Study 1

The Series Learning Potential Test was administered three times to groups of bright, dull-to-average, and subnormal (EMR) children in the middle, elementary school years. A group training session was interpolated following the second test administration. Three scores were calculated for each subject: practice score ($T_2 - T_1$), training score ($T_3 - T_2$), and a combined gain score ($T_3 - T_1$).

Method

Subjects. Subjects were 126 white children (58 boys, 68 girls) in the third, fourth, and fifth grades of several New England schools. The subjects were divided into three groups according to their IQ scores on the Test of General Ability (Flanagan, 1960):

bright normal [n = 64, 21 boys and 43 girls; M IQ = 113 (±12); IQ ⩾100; predominantly from middle-class, suburban homes];

dull-to-average [n = 37, 17 boys and 20 girls; M IQ = 85 (±7); 80 ⩽ IQ ⩽ 99; predominantly from blue-collar homes in an inner-city district]; and

subnormal (EMR), drawn from special classes for the mentally retarded [n = 25, 20 boys and 5 girls; M IQ = 68 (±7); IQ ⩽ 80; from blue-collar homes in an inner-city district].

There was no indication in the school records of organic brain pathology in any subject.

Instruments. In this study, the Test of General Ability (Flanagan, 1960) was used to divide the subjects into three IQ groups, while the three learning-potential scores (Babad & Budoff, 1971) served as the dependent variables.

The Test of General Ability is a group IQ test, consisting of verbal and reasoning parts. Both parts are multiple-choice tests, with pictorial stimuli in the verbal part and abstract symbols in the reasoning part. Three scores are derived from the Test of General Ability: Verbal IQ, Reasoning IQ, and Total IQ.

The Series Learning Potential Test, as described above, consists of two, equivalent, 65-item multiple-choice tests (alternate form reliability is .84). Form A is used for the pretest and Form B for the posttest. Each form is administered as a power test and usually takes 30–35 minutes to complete. The training booklet consists of 17 items, all of which are solved together by the entire class. Five strategies are taught and practiced during training:

1. The students learn to identify each concept that changes.
2. They learn to "sing the tune" for each concept, one at a time as an organizational or chunking aid to identify the pattern of each concept.
3. To reduce the memory load in a multiconcept item, the student crosses out the wrong choices for each concept.
4. The child learns to reverse the direction in which the tune is sung when the location of the blank space calls for such reversal, for example, blank at the beginning of a series.
5. The child learns to identify the starting point of a concept when the series starts in the middle of the pattern.

All 17 problems are successfully solved by the end of the 30–45 minute training session.

Procedure. The study was conducted during four sessions in the spring of the school year. The Series Learning Potential Test was first administered in three sessions, with two-day intervals between sessions. All subjects received Form A (pretest) in the first session. In the second session, half of the subjects received Form A again, and the other half, Form B. For both groups, the training session followed immediately. In the third session, all subjects received Form B (posttest). The Test of General Ability was administered several days later. All tests were administered by experienced examiners, following standard instructions. In each class, all tests were administered by the same person.

Results

Practice, training, and combined gain scores were calculated for the three groups. The patterns of these scores are presented in Figure 1. The means and standard deviations of the three groups in initial and final testing are presented in Table 1.

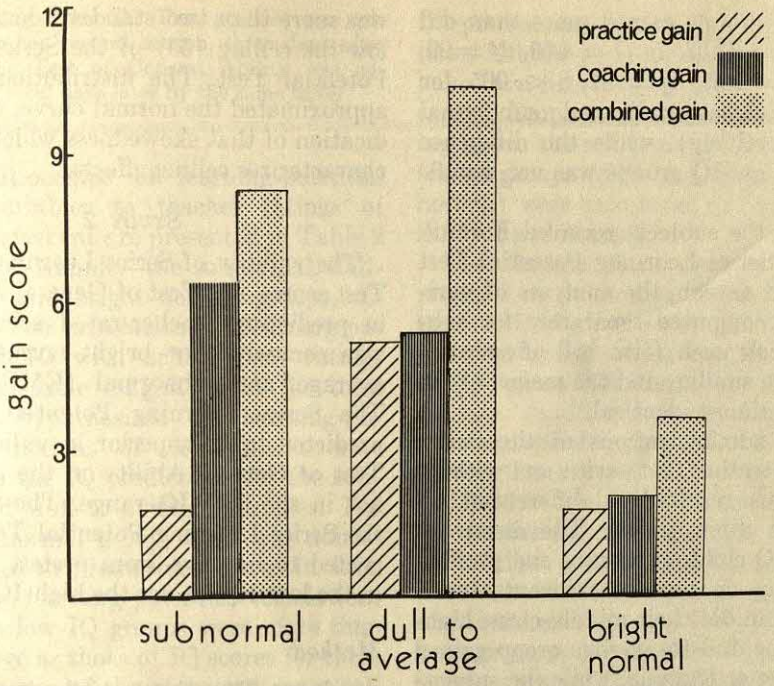


FIGURE 1. Practice gain, coaching gain, and combined gain of the three groups.

To test the first and second hypotheses, separate one-way analyses of variance were computed on practice, training, and combined gain effects for the three IQ groups. These analyses were followed by t tests between the three possible pairs of groups. As can be seen in Figure 1 and Table 1, the first and second hypotheses were supported. Repeated administrations without training resulted in differential score increments for the three groups ($F = 16.3$, $df = 2/125$, $p < .001$). The dull-to-average group gained from practice significantly more than did the subnormal ($t = 2.54$, $df = 60$, $p < .01$) and the bright–normal groups ($t = 3.33$, $df = 99$, $p < .001$), while the difference in practice gains between the latter groups was nonsignificant. Training produced differential increments in score ($F = 2.4$, $df = 2/125$, $p < .10$), but the pattern differed from that found for practice. Both the subnormal and the dull-to-average groups gained more from training than did the bright–normal group ($t = 2.69$, $df = 87$, $p < .005$; $t = 2.26$, $df = 99$, $p < .025$, respectively). The highest mean gain in training was attained by the subnormal group, but the difference in the training effect between the two low-IQ groups was not significant.

The analysis of variance of the combined gain scores indicated differential increments in score ($F = 9.32$, $df = 2/125$, $p < .001$).

TABLE 1

MEANS AND STANDARD DEVIATIONS OF SLPT AND TOGA SCORES OF THE THREE GROUPS IN STUDY 1

Group	SLPT 1, initial performance			SLPT 2, final performance		TOGA IQ	
	n	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Bright–normal	64	51.9	5.0	55.4	4.7	113	12.0
Dull-to-average	37	37.6	10.1	47.3	9.9	85	7.0
Subnormal	25	26.4	10.7	35.0	12.3	68	7.0

Note. Abbreviations: SLPT = Series Learning Potential Test and TOGA = Test of General Ability.

Both low-IQ groups gained more than did the bright–normal group ($t = 4.03$, $df = 99$, $p < .001$; $t = 2.89$, $df = 87$, $p < .005$, for the bright versus dull and subnormal groups, respectively), while the difference between the low-IQ groups was not significant.

Since half the subjects received Form A or B of the Series Learning Potential Test in the second session, the analyses of variance were recomputed separately for subjects who took each form. All effects and patterns were similar, and the means of the groups were almost identical.

Repeated administrations of the Series Learning Potential Test—with and without training—thus resulted in differential effects on the three groups. The disadvantaged, low-IQ children learned and profited from training on problem-relevant strategies more than did their middle-class, high-IQ peers. The dull-to-average group gained from practice or training, while the subnormal sample seemed to profit most from training.

The relative spread of Test of General Ability and Series Learning Potential Test scores for the three groups is also indicative of the sensitivity of learning-potential assessment in the low-socioeconomic-status, low-IQ groups. While the standard deviation of IQ scores for the bright–normal group (12) is almost twice as large as the standard deviations for both low-IQ groups (both 7 points), the picture is reversed for Series Learning Potential Test scores. The standard deviations of both low-IQ groups are twice as large as those of the bright–normal group, both in initial and final testing (Table 1). The differential effect of training is indicated in the changes of the Series Learning Potential Test standard deviations from initial to final testing. A slight shrinkage in standard deviation was found for the bright group as compared with the 20% increase in the standard deviation of the subnormal group.

The possibility that the observed pattern of results was caused by a ceiling effect for the bright–normal group in the posttest was explored by checking this group's performance in the posttest. The mean score (55.4) was more than two standard deviations below the ceiling (65) of the Series Learning Potential Test. The distribution of scores approximated the normal curve, with no indication of that skewedness which typically characterizes ceiling effects.

Study 2

The validity of Series Learning Potential Test scores and Test of General Ability IQs in predicting teacher-rated school success was compared for bright–normal, dull-to-average, and subnormal (EMR) samples. The Series Learning Potential Test was predicted to be superior in validity to the Test of General Ability in the low-IQ but not in the high-IQ range. The validity of the Series Learning Potential Test was expected to improve from pretest to posttest in the low-IQ but not the high-IQ range.

Method

Subjects. The subjects of the first study were selected from separate schools, the EMR sample consisting of students in segregated special classes for the retarded. To equate samples, schools, classes, and teacher grades, a new sample was selected with a wide representative range of ability in each class. In this sample, all EMR subjects were integrated in regular classes.

Subjects were 207 white children in 12 third-grade classes of three elementary schools. All schools were located in a small New England town, with a predominantly white, working-class population. The subjects were divided into three groups according to their Test of General Ability IQ scores: bright–normal [$n = 76$, $M = 113$ (± 9), IQ $\geqslant$ 100]; dull-to-average [$n = 95$, $M = 88$ (± 6), $80 \leqslant$ IQ $\leqslant 99$]; and subnormal [$n = 36$, $M = 72$ (± 6.5), IQ $\leqslant 80$]. There was no indication in the school records of organic brain pathology in any subject. Subjects for all three groups were found in each of the 12 classes.

Instruments. As in the first study, the Series Learning Potential Test and the Test of General Ability were used. School achievement was measured by teacher ratings for reading, spelling, arithmetic, "general achievement," and "academic potential." The ratings were done on an 11-point scale, corresponding to the letter grade system which the teachers are accustomed to use in that school system.

Procedure. The study was conducted in three sessions during the spring of the school year. In the first session, the Series Learning Potential Test pretest (Form A) was administered to all students, followed by the standardized training. The teachers completed the rating scales during this period. The

Series Learning Potential Test posttest was administered in the second session, three days after the first. The Test of General Ability was administered several days later by the same tester.

Results

The relationships of learning-potential and IQ variables to teacher ratings of school achievement are presented in Table 2 for the entire sample—the subnormal, dull-to-average, and bright–normal groups, respectively. The correlations of learning potential and IQ with achievement ratings were of the same magnitude in the total sample. As hypothesized, the learning-potential predictions of achievement were higher than the IQ predictions for the subnormal and dull-to-average samples. For the bright–normal group, there was almost no difference in predictive power. In fact, the posttest, learning-potential correlations for the two low-IQ groups were more than twice as large as those of IQ scores for these groups. Because of the restricted range, all subsample correlations decreased; but the predictability of IQ scores suffered more from this restriction than did the learning-potential scores for the low-IQ groups. The third hypothesis was thus supported. The validity of the Series Learning Potential Test was not inferior to that of the Test of General Ability IQ in the overall sample and was superior to IQ in the restricted low-IQ range.

As to the fourth hypothesis, the expected changes in validity of the Series Learning Potential Test from pretest to posttest were found (increase for the low-IQ groups; no increase for the high-IQ group), but the changes were small and inconclusive. The hypothesis was thus not confirmed by the data. One should also note the relative superiority of Reasoning-IQ predictions over Verbal-IQ predictions for both low-IQ groups and the absence of such differences in the bright–normal group.

One may view another strength of the training-based learning-potential measurement by examining the proportion of IQ-defined EMR subjects who attain the mean pretest score of their nonretarded agemates following training. According to the trainability hypothesis, EMR subjects who reach the pretraining, reasoning level of nonretardates following a short training session may have been falsely identified as mentally retarded. Proportions of EMR subjects falling above the mean pretest score of nonretarded groups (dull-to-average and bright–normal) were calculated for the samples in Study 1 and Study 2.

The changes from pre- to posttraining proportions are quite marked, indicating that following a short problem-relevant learning experience, substantial proportions of IQ-defined EMR subjects do attain the average (noncoached) reasoning level of their nonretarded peers. The fifth hypothesis was thus supported. Among special-class EMRs, 36% and 13% attained scores above the pretraining means of the dull-average and bright–normal groups, respectively. The mean EMR posttest performance exceeded the mean pretest level of the dull-to-average group, despite the 16-point IQ difference between the groups, and 61% of the EMRs scored above this level. More than a third of the EMR subjects (36%) surpassed the mean pretest score of the bright group, despite a 41-point (!) difference between the average IQs of the two groups.

Discussion

The results indicate that learning-potential testing did show a higher level of differentiation among low-IQ, low-socioeconomic-status subjects than the standard test of intelligence. Utilization of a learning-potential assessment paradigm confirmed considerable ability to reason among low-IQ children which was not indicated by standard measures of intelligence. Disadvantaged dull-to-average and EMR groups profited from training on problem-relevant strategies more than did the bright–normal group, while the dull-to-average group gained from practice or training appropriate to the task. Substantial proportions of IQ-defined EMR children reached the reasoning level of their brighter peers following a short training session on problem-relevant strategies. The observed sensitivity of learning-potential measurement in the low-IQ range was further borne out by the superiority of the Series Learning Potential

Test to the Test of General Ability IQ in predicting school achievement. While the predictive power of both tests was of the same magnitude in the bright–normal group, the posttraining score was a better predictor than IQ in the dull-to-average and EMR groups.

To permit reasonable predictions concerning educability in a school setting, reasoning abilities of the child must be tested. These abilities are inevitably tainted by differences in the prior cultural experiences of the subjects. By providing all children with directed training on problem-relevant strategies, the learning-potential paradigm seeks to bring all children to a more equalized starting point prior to testing, thereby reducing the advantage of the high-socioeconomic-status children. Indeed, there was a considerably greater overlap of scores among groups differing in IQ following training than there had been prior to training. Given the opportunity to become familiar with the demands of the task and to learn how to approach the problems, the lower-IQ children from blue-collar and poor backgrounds displayed considerably more competence, confidence, and sense of challenge on the task than they had previously. In Hunt's (1961) terms, it could be said that for the posttraining measurement, there was a better match between the children's existing schemata and the demands of the situation. For the middle-class children, the "problem of the match" did not exist, and they performed as well on the pretest as on the posttest.

The increased overlap between reasoning curves of high- and low-socioeconomic-status subjects following coaching indicates that the Level II ability of portions of disadvantaged populations is greater than that posited by Jensen on the basis of one-shot reasoning tests. Training permitted identification of those poor, low-IQ children who have potential for functioning within a framework of Level II thinking.

Budoff, Meskin, and Harrison (1971) have demonstrated how to apply this evidence of reasoning ability within a classroom curriculum. They tested the acquisition of principles of electricity in a manipulative science program with groups of special and regular class students. They found that learning-potential status best differentiated levels of attainment following this course, while neither IQ nor special versus regular class placement could distinguish levels of achievement after the course.

Judging by the success of the learning-potential approach, all successful teaching programs would seem to profit by incorporating the following: the initial utilization of channels in which children are less deficient, the provision of the necessary preparatory learning experiences, and the crea-

TABLE 2

COEFFICIENTS OF CORRELATION OF SLPT AND TOGA SCORES WITH TEACHER RATINGS

Teacher ratings	Predictor																				
	Total sample (N = 207)					Subnormal sample (n = 36)					Dull-to-average sample (n = 95)					Bright–normal sample (n = 76)				
	SLPT		TOGA			SLPT		TOGA			SLPT		TOGA			SLPT		TOGA		
	Pretest	Posttest	Verbal	Reasoning	IQ	Pretest	Posttest	Verbal	Reasoning	IQ	Pretest	Posttest	Verbal	Reasoning	IQ	Pretest	Posttest	Verbal	Reasoning	IQ
Reading	.50	.50	.45	.49	.54	.24	.28	−.01	.20	.12	.35	.35	.05	.15	.09	.35	.33	.27	.23	.29
Spelling	.45	.49	.45	.47	.53	.26	.38	.19	.20	.20	.23	.26	.08	.17	.17	.32	.32	.26	.22	.33
Arithmetic	.49	.53	.46	.50	.52	.35	.49	.18	.33	.18	.27	.29	.14	.16	.18	.43	.40	.26	.34	.32
General achievement	.51	.54	.49	.52	.56	.33	.40	.07	.25	.15	.31	.34	.08	.22	.19	.40	.38	.33	.26	.32
Academic potential	.49	.50	.43	.47	.50	.27	.22	.11	.01	.13	.35	.44	.04	.23	.11	.37	.35	.25	.25	.29
Average correlation	.49	.51	.46	.49	.53	.29	.35	.11	.20	.16	.30	.34	.08	.19	.15	.37	.35	.27	.25	.31

Note. Abbreviations: SLPT = Series Learning Potential Test and TOGA = Test of General Ability.

tion of subjective feelings of mastery and success.

REFERENCES

BABAD, E. Y., & BUDOFF, M. A preliminary manual of the Series Learning Potential Test. Unpublished manuscript, Research Institute for Educational Problems, 1971.

BUDOFF, M. Learning potential: A supplementary procedure for assessing the ability to reason. *Seminars in Psychiatry,* 1969, **1,** 278–290.

BUDOFFF, M., & CORMAN, L. Demographic and psychometric factors related to improved performance on the Kohs learning potential procedure. *American Journal of Mental Deficiency,* 1974, **78,** 578–585.

BUDOFF, M., CORMAN, L., & LITZINGER, S. A manual for the Raven's Learning Potential Test. *Studies in Learning Potential,* 1973, **3,** No. 59.

BUDOFF, M., MESKIN, J., & HARRISON, R. H. An educational test of the learning potential hypothesis. *American Journal of Mental Deficiency,* 1971, **76,** 159–169.

EELLS, K., DAVIS, A., HAVIGHURST, R. J., HERRICK, V. E., & TYLER, R. *Intelligence and cultural differences.* Chicago: University of Chicago Press, 1951.

FLANAGAN, J. C. *Test of General Ability: Technical report.* Chicago: Science Research Associates, 1960.

HUNT, J. MCV. *Intelligence and experience.* New York: Ronald Press, 1961.

JENSEN, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review,* 1969, **39,** 1–123.

ORTAR, G. Improving test validity by coaching. *Educational Research,* 1960, **2,** 137–142.

ZIGLER, E. Mental retardation: Current issues and approaches. In L. W. Hoffman & M. L. Hoffman (Eds.), *Review of child development research.* Vol. 2. New York: Russell Sage Foundation, 1966.

(Received August 17, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 3, 448-456

INSTRUCTIONAL OBJECTIVES AS DIRECTIONS TO LEARNERS:

EFFECT OF PASSAGE LENGTH AND AMOUNT OF OBJECTIVE-RELEVANT CONTENT[1]

ROBERT KAPLAN[2]
Bell Laboratories, New Brunswick, New Jersey

ERNST Z. ROTHKOPF
Bell Laboratories, Murray Hill, New Jersey

The effects of number of objective-relevant sentences on learning from texts ranging in length from about 800 to 3,000 words were investigated in two related experiments. The main findings were (*a*) the use of objectives as directions resulted in increased performance; (*b*) the likelihood of mastery of any objective (intentional learning) decreased with the number of objective-relevant sentences but was not related to passage length; (*c*) incidental learning decreased with passage length; (*d*) study time increased with the number of objectives and with passage length; and (*e*) specifically stated objectives resulted in greater intentional learning than generally stated objectives.

The use of statements of explicit goals as directions to learn from text materials has been investigated in a study by Rothkopf and Kaplan (1972). They found that explicit objectives resulted in better learning performance than did directions to learn as much as possible, both on test items that were relevant to the objectives and on incidental test items. Specifically described objectives were found to produce higher performance than more generally stated objectives. Further, the likelihood of mastering any given objective was found to decline with increases in the density of objective-relevant sentences in the text.

The present experiments had three main purposes. First, an attempt was made to replicate the previous experiment with an extended range of text lengths and densities of objective-relevant text components. The second purpose was to clarify the effective elements of the density variable. This factor, which was defined in the previous study as the ratio of sentences judged to be objective relevant to the total number of sentences in the text, actually involves several distinguishable but correlated variables. These include (*a*) the number of relevant sentences in the text, (*b*) the ratio of relevant sentences to the total number of sentences in the text, and (*c*) the number of sentences in the direction specifying these objectives to the subject. The purpose of the two studies reported here was to determine whether Variables *a* and *b* influenced the the likelihood of mastering any given objective. Another important goal was to determine whether the specificity and number of objectives affected the amount of time used for studying the text.

Experiment 1

Method

Materials

Passages. Three experimental passages used in a previous study (Rothkopf & Kaplan, 1972) were adapted for use in this experiment. The passages were selected from three textbooks prepared by the Systems Training Department, Bell Telephone Laboratories. These basic passages presented information about printing specifications for designing forms (Passage 1) and an introduction to busi-

[1] We are indebted to E. M. Burgin and D. M. Lynch for their valuable assistance in helping throughout all phases of the experiments.

[2] Requests for reprints should be sent to Robert Kaplan, Bell Laboratories, CP1 L101, P.O. Box 2020, New Brunswick, New Jersey 08903.

ness information systems (Passage 2) and total system development (Passage 3).[3] The passages comprised 842, 1,091, and 1,120 words and 60, 54, and 55 sentences, respectively.

Passage length. The three basic passages were used singly and in combinations to form eight experimental texts, 56, 113, and 169 sentences in length. Length 56 was achieved by using separately each of the three basic passages ($\bar{X}$ length = 56 sentences). The second passage, Length 113, was achieved by combining two Length 56 passages. Three such two-passage combinations were used: Passages 2, 1; 2, 3; and 3, 1 ($\bar{X}$ length = 113 sentences). Two versions of the longest experimental passage, Length 169, were constructed by combining the three basic passages in order 2, 3, 1 and 2, 1, 3 ($\bar{X}$ length = 169 sentences).

Objectives. The objectives were the same as those used by Rothkopf and Kaplan (1972). An objective was defined as a direction to the subject to learn a particular instructional point in the passage. Two types of objectives were prepared. First, a set of specifically phrased objectives (relevant to one text sentence) was written for each Length 56 passage. Second, a matching set of generally phrased objectives (relevant to clusters of 2–5 adjacent text sentences) was prepared for each Length 56 passage. The relevance of passage sentences to objectives was empirically determined in a preliminary study (for details see Rothkopf & Kaplan, 1972). The match between 1 general objective and 2–5 specific objectives was based on their sharing the exact same relevant sentences in the passage.[4]

Four sets of specific and general objectives were selected such that their relevant sentences in the passage resulted in four densities (40%, 60%, 75%, and 85%) for all passage lengths. This designation of the four density levels refers to the approximate proportion of objective-relevant sentences in the passage. All sets of objectives were about evenly distributed throughout the passage. Increases in density for a given passage length were produced by nesting; that is, higher densities were produced by adding to the objectives used in the lower densities. The mean number of relevant sentences for the four densities of Length 56 was 22, 34, 41, and 48. It should be noted that only the first group of 22 sentences, which comprised the 40% density, were common to or reproduced across all densities (common intentional). No objectives were written for those sentences that remained after the sentences of 85% density were selected. These remaining sentences provided a measure of incidental learning that was common to every density level (common incidental). The density levels for Length 113 and Length 169 were held constant by combining the objectives for the respective Length 56 passages. The number of objectives and the total number of sentences varied across passage length, but the proportion of objectives to total sentences (density) remained the same.

Tests. Completion-type test questions were constructed by removing one or two substantive words from an approximate paraphrasing of passage sentences and replacing word(s) with a line of uniform length. A test question was achieved for almost every sentence in the basic Length 56 passages. This meant that there was a test question for every sentence relevant to an objective (intentional items) and for most of the nonobjective sentences (incidental items) for each treatment. Combinations of these Length 56 test questions were used for testing Length 113 and Length 169.

Procedure

Groups of approximately 100 subjects were tested in each experimental session. Group procedures, following those of Rothkopf and Coke (1968), were used. Each subject received a package of materials consisting of three sequentially numbered manila envelopes that subjects completed in the numbered order. The three numbered envelopes contained, respectively, (*a*) a set of objectives and a passage, (*b*) a test on the assigned passage, and (*c*) a nonexperimental task to occupy the subject while the other subjects completed their experimental task. For the reference groups, Envelope 1 contained a passage and conventional directions to learn everything in the passage. The experimental subjects were informed that they would be tested only on the information in the passage that was relevant to the list of objectives. However, they were tested on almost every sentence in the passage. This procedure permitted testing of incidental (no objective) as well as intentional (with objectives) learning. The list of objectives was available to the subjects while they read the passage. The objectives were listed in the same sequence that the corresponding relevant sentences occurred in the passage.

Eighteen subjects were assigned to each of 24 treatments (n = 432). Fifty-four additional subjects comprised three reference groups, one for each passage length (n = 18 per length). These groups

[3] We are grateful to F. L. Stevenson, Head, Systems Training Department, Bell Telephone Laboratories, New Brunswick, New Jersey, for permitting us to use the experimental material.

[4] An example of a cluster of three specific objectives was, "Learn about the physical appearance of Gothic type. Learn about the physical appearance of Italic type. Learn about the physical appearance of Roman type." The matching general objective for this cluster of specific sentences was, "Learn about the physical appearance of the three kinds of type discussed." Both the general and the specific objectives were relevant to the following three adjacent text sentences: "Gothic type has no serifs or doodads to confuse the reader. Conversely, Italic type does have serifs and a distinctive slant causes it to stand out. The physical appearance of Roman type is similar to Italic in that it has serifs, but Roman type does not slant."

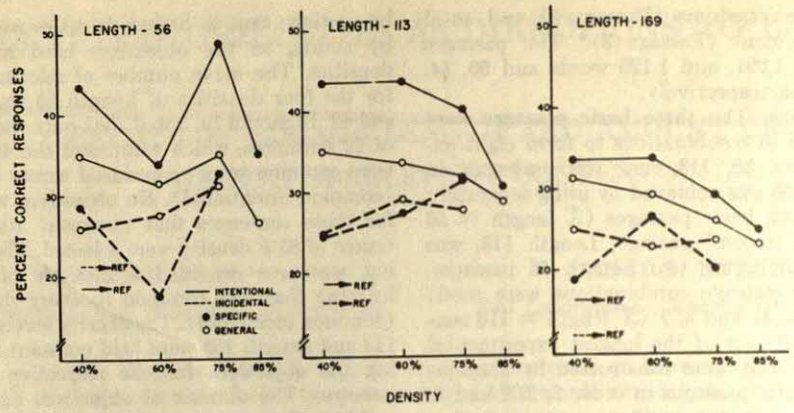


FIGURE 1. Mean proportion of correct responses for intentional and incidental test items that were common to all treatments as a function of passage length, density, and specificity of objectives. (Reference [REF] groups that received no objectives are shown for each passage length.)

read the passages with directions to learn everything in the passage. This treatment was similar to conventional directions in learning experiments which do not usually present objectives but give broad directions to learn everything possible.

Subjects

Students from the following seven New Jersey high schools served as paid volunteers in the experiment: Battin High School, Elizabeth; Bernards High School, Bernardsville; Caldwell High School, West Caldwell; Arthur L. Johnson Regional High School, Clark; Livingston High School, Livingston; Somerville High School, Somerville; and Westfield High School, Westfield.[5] They consisted of 177 males and 309 females between the ages of 13 and 18 (n = 486). The experimental sessions were conducted in the high school cafeterias after the last school period.

Results

Intentional Learning

The objective-relevant test items that were common to all treatments were analysed with a 2 × 3 × 4 analysis of variance containing (a) two levels of objective specificity (specific and general); (b) three levels of passage length (56, 113, and 169 sentences); and (c) four levels of density (40%, 60%, 75%, 85%).[6] Arc sine transformations were used on the proportion of correct responses. Treatment means are shown in Figure 1. The analysis supports the following conclusions.

The specifically stated objectives ($\bar{X}$ = .37) produced higher intentional learning scores than did the generally stated objectives ($\bar{X}$ = .31; F = 8.61, df = 1/408, p < .01). As seen in Figure 1, the likelihood of mastering any single intentional item generally decreased with passage length ($\bar{X}$ Lengths 56, 113, and 169 = .36, .36, and .29, respectively; F = 4.69, df = 2/408, p < .01). The likelihood that any given intentional item was correctly recalled tended to decline with increases in density, although the 75% density condition for Passage Length 56 appeared to be aberrant in this respect ($\bar{X}$ Densities 40%, 60%, 75%, and 85% = .37, .34, .35, and .29, respectively; F = 2.89, df = 3/408, p < .05). The density effect was somewhat more pronounced for the specific than the general objectives treatments. None of the interactions reached significance.

Incidental Learning

The nonobjective-relevant test items that were common to all treatments were analysed with a 2 × 3 × 3 analysis of variance.

[5] We extend special appreciation to the following high school principals for recruiting student participants and permitting the use of their facilities: D. Parisi, T. Froisland, N. Bussiere, R. Hough, L. Hurley, M. Crisci, and A. Bobal, respective to the listed schools.

[6] Appreciation is extended to J. R. Gossman for his participation in the analysis of data.

The first two factors were the same as those for intentional learning. However, the third factor (density) was reduced to three levels (40%, 60%, and 75%) in order to increase the number of incidental learning test items that were common to all densities. That is, there were only four items common to all conditions which were available to measure Length 56 incidental learning when 85% density was included in the analysis. The elimination of 85% density permitted 12 items common to all conditions to be used as a measure of incidental learning. Arc sine transformations were again used on the proportion of correct responses. The incidental learning means are given in Figure 1.

Specificity of objectives and density of objective-relevant sentences in the passage had little or no effect on incidental learning ($F = .42$, $df = 1/306$ and $F = .65$, $df = 2/306$, respectively). The passage length main effect was significant ($\bar{X}$ Lengths 56, 113, and 169 = .27, .28, and .23, respectively; $F = 3.21$, $df = 2/306$, $p < .05$). The prevailing trend was that the likelihood of correct response to any given incidental item decreased as a function of passage length. None of the interactions were significant. As expected, performance on incidental items ($\bar{X} = .26$) was poorer than on intentional items ($\bar{X} = .34$).

Reference Group Comparisons

The reference treatment resulted in weaker performance with respect to both intentional and incidental test items than did the experimental treatment. The only exception was the incidental learning score for the 60% density/specific objectives condition in the Length 56 passages. All comparisons between reference and treatment groups were made with t tests. Figure 1 shows that intentional learning scores for the treatment groups were higher than the reference group scores (Length 56, $t = 4.13$, $df = 160$, $p < .001$; Length 113, $t = 4.65$, $df = 160$, $p < .001$; Length 169, $t = 3.90$, $df = 160$, $p < .001$). In addition, t tests were used to compare treatment with reference groups for incidental learning. Length 56 incidental learning scores were not significantly different from the reference group scores ($t = 1.28$, $df = 160$, $p < .05$). However, significantly higher incidental learning was found for the Length 113 treatment group over the reference group ($t = 2.23$, $df = 160$, $p < .05$) and for the Length 169 treatment group over the reference group ($t = 1.79$, $df = 160$, $p < .05$). Similar to the incidental learning scores of the experimental treatments, the reference group scores generally decreased as a function of passage length ($\bar{X}$ Lengths 56, 113, and 169 = .20, .19, and .16, respectively). This decrease was not statistically reliable.

Discussion

The results of Experiment 1 replicated all of the major findings reported by Rothkopf and Kaplan (1972) with passages up to three times as long as those in the original study and with a much greater range of densities. The new result in Experiment 1 was that both intentional and incidental learning were observed to decrease with passage length. It should be noted, however, that it is impossible to determine from Experiment 1 whether this finding with respect to intentional learning was due to passage length or the greater number of objectives in the longer passages. This was due to the fact that experimental densities of objective-relevant sentences were held constant for all three passage lengths. Consequently, the average absolute number of objective-relevant sentences was greater for the longer than for the shorter passages.

Experiment 2

The primary purpose of Experiment 2 was to determine whether the decremental effects on intentional learning were due to increases in density or the absolute number of objective-relevant sentences. This was accomplished by observing learning performance associated with various passage lengths when the number of objective-relevant sentences was the same for each passage length but density was permitted to vary. Under these experimental arrangements, it was possible to (*a*) evaluate the effect of passage length with the number of

objectives held constant and (*b*) evaluate the effect of a given number of objectives for various passage lengths. A density effect on learning would be indicated if, for a given number of objectives, performance was observed to be stronger for longer passages. On the other hand, decrements due to an increasing number of objectives but constant levels of performance for any given number of objectives across various passage lengths would indicate that increasing the number of objectives was responsible for the "density" decrements observed in Rothkopf and Kaplan (1972) and in Experiment 1. Another purpose of Experiment 2 was to obtain inspection time data for each treatment.

Method

Design and Procedure

The passages and basic objectives of Experiment 1 were used here. The data collected from six treatments in Experiment 1 were combined with five new experimental groups to produce an experimental design that involved three passage lengths (56, 113, and 169 sentences), each of which was used with objectives referring to 22, 34, 44, and 100 relevant sentences (except, of course, the impossible combination of Passage Length 56 with 100 relevant sentences). All combinations of passage length and number of relevant sentences were used with both specifically and generally phrased objectives. This led to an incomplete factorial design involving three passage lengths, four levels of relevant sentences, and two degrees of specificity in the phrasing of objectives for a total of 22 experimental treatments. Three reference groups, one for each passage length and identical to those in Experiment 1, were also used in the analysis. The procedure for administering the experimental materials and the test questions used in Experiment 2 were the same as those used in Experiment 1.

Three dependent measures were used: performance on (*a*) intentional and (*b*) incidental test items common to all treatments and (*c*) passage inspection time. The use of Measures (*a*) and (*b*) was described in Experiment 1. The third measure, inspection time, was obtained by requiring the subject to record the start and stop reading times using a large digital clock. The experimental effects on these three measures were evaluated with two separate analyses because of the missing cell in the incomplete factorial design.

Subjects

The data from 216 students who participated in Experiment 1 were used here. In addition, 180 students selected from the same high schools and participating simultaneously with those in Experiment 1 completed the number of subjects required for the treatment groups of this experiment. The same 54 students served as a reference group in both experiments. Altogether, there were 450 subjects consisting of 158 males and 292 females. This permitted the assignment of 18 subjects to each of the 22 treatments and the 3 reference groups.

Results

Common Intentional and Incidental Test Items

Two analyses were performed for intentional and incidental items that were common to all treatments. The first was a $2 \times 3 \times 3 \times 2$ analysis of variance with (*a*) two levels of objectives (general and specific); (*b*) three levels of relevant sentences (22, 34, and 44 sentences); (*c*) three levels of passage length (56, 113, and 169 sentences); and (*d*) two levels of learning (intentional and incidental) with repeated measures on the last factor. Arc sine transformations were used for proportion scores. The means for these data are shown in Figure 2 and support the conclusions below.

In general, larger numbers of relevant sentences resulted in a smaller likelihood of successful performance on any given intentional item except for the aberrant reversal between the 34- and 44-sentence conditions. The effect due to number of relevant sentences was significant ($F = 4.01$, $df = 2/306$, $p < .05$). Figure 2 also indicates that performance on intentional items for any given number of relevant sentences appears relatively constant across various passage lengths.

By holding the number of relevant sentences constant across the various passage lengths, the passage length effect on intentional learning was removed ($\bar{X}$ Lengths 56, 113, and 169 = .38, .39, and .38, respectively). Paired comparisons among the three conditions were made with Newman-Keuls tests. The largest difference between Length 56 and Length 113 yielded a q of 1.29 ($df = 306$, $r = 2$, $p < .05$). This suggests that the passage length effect on intentional learning observed in Experiment 1 (see Figure 1) was due to a correlation between density

and passage length; that is, when passages of various lengths were matched with respect to density, as in Experiment 1, the number of relevant sentences was greater for the longer passages.

The difference between general ($\bar{X} = .29$) and specific ($\bar{X} = .34$) objectives for intentional and incidental learning was not statistically reliable ($F = 3.36$, $df = 1/306$, $.05 < p < .10$). One interaction that reached significance was between objectives and learning ($F = 14.63$, $df = 1/306$, $p < .001$). This was a consequence of the fact that, for intentional learning, specific objectives ($\bar{X} = .43$) resulted in higher scores than general objectives ($\bar{X} = .34$; $q = 5.00$, $df = 306$, $r = 2$, $p < .001$), while there were no differences between specific ($\bar{X} = .24$) and general ($\bar{X} = .24$) directions on incidental learning ($q = .50$, $df = 306$, $r = 2$).

Performance on intentional items ($\bar{X} = .38$) was better than on incidental items ($\bar{X} = .24$; $F = 89.56$, $df = 1/306$, $p < .001$). There appeared to be no systematic relationship between incidental scores and the number of relevant sentences. Performance on incidental items did, however, tend to decrease with passage length although not significantly ($\bar{X}$ Lengths 56, 113, and 169 = .32, .25, and :16, respectively).

The Number of Relevant Sentences × Passage Length interaction ($F = .79$, $df = 4/306$) was not significant. That is, for a given number of relevant sentences, variations in passage length yielded similar scores. This finding further suggests that the absolute number of intentional sentences, rather than density, influenced performance on intentional items.

Another analysis of variance ($2 \times 4 \times 2 \times 2$) was performed on these data in order to permit the inclusion of 100 objectives in the analysis of Passage Lengths 113 and 169. The factors were (*a*) two levels of objectives (general and specific); (*b*) four levels of relevant sentences (22, 34, 44, and 100 sentences); (*c*) two levels of passage length (113 and 169 sentences); and (*d*) two levels of learning (intentional and incidental). Figure 2 also includes the means for these data. The results were the same as those found in the first analysis except

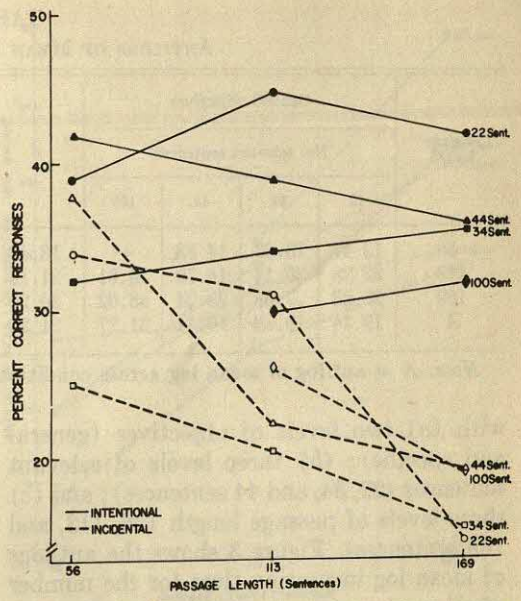


FIGURE 2. Mean proportion of correct responses for intentional and incidental test items that were common to all treatments as a function of passage length. (Data points for treatments with the same number of objective-relevant sentences but different passage lengths are connected; Sent = sentences.)

that here performance with specific objectives ($\bar{X} = .31$) was significantly greater than with general objectives ($\bar{X} = .26$; $F = 7.03$, $df = 1/272$, $p < .01$). The number of relevant sentences main effect was significant ($\bar{X}$ 22, 34, 44, and 100 sentences = .33, .26, .29, and .27, respectively; $F = 3.06$, $df = 3/272$, $p < .001$). The Objectives × Learning interaction was again significant in this analysis ($F = 8.79$, $df = 1/272$, $p < .01$).

Inspection Time

All analyses of inspection time were performed on log transformations of time data in order to reduce the effect of extremely long times. The antilogs of the mean log for time data for each treatment and for the reference groups are presented in Table 1. Two analyses were performed on log transformations of the inspection times. The designs of these analyses were identical to the two test performance analyses previously presented in this experiment.

The first analysis was a $2 \times 3 \times 3$ design

TABLE 1

ANTILOGS OF MEAN LOG INSPECTION TIME

Passage length	General objectives				A	Specific objectives				A	Reference groups
	No. relevant sentences					No. relevant sentences					
	22	34	44	100		22	34	44	100		
56	12.47	13.03	14.83	—	13.40	17.14	16.98	25.88	—	19.63	8.79
113	22.28	20.14	16.79	26.61	21.19	23.77	29.38	25.47	25.53	25.94	18.83
169	25.23	29.58	29.24	38.02	30.20	25.53	32.66	35.32	39.63	32.89	24.14
A	19.14	19.82	19.41	31.77	21.26	21.83	25.35	28.58	31.77	26.21	15.87

Note. *A* = antilog of mean log across conditions.

with (*a*) two levels of objectives (general and specific); (*b*) three levels of relevant sentences (22, 34, and 44 sentences); and (*c*) three levels of passage length (56, 113, and 169 sentences). Figure 3 shows the antilogs of mean log inspection time for the number of relevant sentences condition as a function of passage length collapsed across general and specific objectives. The results obtained are reported below.

More time was spent inspecting passages with specific objectives than general objectives ($F = 25.29$, $df = 1/270$, $p < .001$). Inspection time increased as a function of the number of relevant sentences, but this effect did not quite reach significance ($F = 2.63$, $df = 2/270$, $.08 > p > .07$). Reading time increased as a function of passage length ($F = 45.80$, $df = 2/270$, $p < .001$). Newman-Keuls analyses showed Length 169 > Length 113 > Length 56 ($q = 5.95$ and 7.63, $df = 220$, $r = 2$, $p < .01$, respectively). The only interaction to attain significance was Relevant Sentences × Passage Length ($F = 2.99$, $df = 4/270$, $p < .05$). This significant interaction was primarily a result of the 44-sentence condition that did not increase between Lengths 56 and 113.

The second inspection time analysis was a 2 × 4 × 2 design with (*a*) two levels of objectives; (*b*) four levels of relevant sentences (22, 34, 44, and 100 sentences); and (*c*) two levels of passage length (113 and 169 sentences). The results of this analysis follow (see Figure 3).

Passages with specific objectives again required more reading time than did those with general objectives ($F = 7.66$, $df = 1/240$, $p < .01$). In this analysis, the number of relevant sentences main effect was significant ($F = 5.02$, $df = 3/240$, $p < .01$). Newman-Keuls analyses showed that 100 relevant sentences required more time than 22, 34, and 44 relevant sentences ($q = 5.17$, $df = 240$, $r = 4$, $p < .01$; $q = 2.74$, $df = 240$, $r = 3$, $p < .01$; $q = 3.95$, $df = 240$, $r = 2$, $p < .01$, respectively). In addition, 34 relevant sentences required more reading time than 22 relevant sentences ($q = 2.43$, $df = 240$, $r = 2$, $p < .01$). Again, the larger Length 169 passage required more reading time than the Length 113 passage ($F = 31.89$, $df = 1/240$, $p < .001$). None of the interactions reached significance in this analysis.

Figure 3 indicates that inspection time for every treatment exceeded the inspection time for the corresponding reference group. Inspection time increased approximately linearly as the experimental passage became longer and the number of relevant sentences for the treatment groups became more numerous.

DISCUSSION

The present experiments confirmed previous observations (Rothkopf & Kaplan, 1972) that the likelihood of attaining any given objective was reduced as the number of relevant sentences in the passage increased, and provided important evidence about the source of this effect. Experiment 2 indicates that the observed decline in performance was due to increases in the absolute number of objective-relevant sentences

in the text and was relatively independent of passage length. The density of relevant sentences in a passage is therefore not an appropriate characterization of the experimental factor that produced decreases in the probability of correct performance on any given objective as a function of the increase in the number of relevant sentences. This characterization was tentatively used in the Rothkopf and Kaplan (1972) study in which only one passage length was used. Experiment 2, in which both the number of relevant sentences and passage length were systematically varied in a factorial design, showed that performance on any given objective-relevant item was relatively constant for any given number of relevant sentences across passages of different lengths. It would not be reasonable, however, to extend these conclusions beyond the number of relevant sentences and passage lengths used in the present experiment. It does not seem plausible that the likelihood of mastering any given objective would be the same in an extremely short and an extremely long experimental passage. The passages used in the present experiment are, however, within the range of lengths customary in school reading assignments.

The relationship between the likelihood of correct performance on any given objective and passage length, observed in Experiment 1, can be accounted for in terms of the performance constancy associated with number of relevant sentences. In Experiment 1, passages of three different lengths but with identical density levels were used Consequently, the average number of relevant sentences was greater for the longer passages even though the passages of various lengths were matched for density. The main effect on intentional learning due to passage length can therefore be understood in terms of the increase in average number of relevant sentences associated with increases in passage length.

The two experiments replicate several of the findings reported by Rothkopf and Kaplan (1972) with passages up to about 1,500 words in length and with sets of objectives that encompass up to 100 sentences in a passage. These findings include

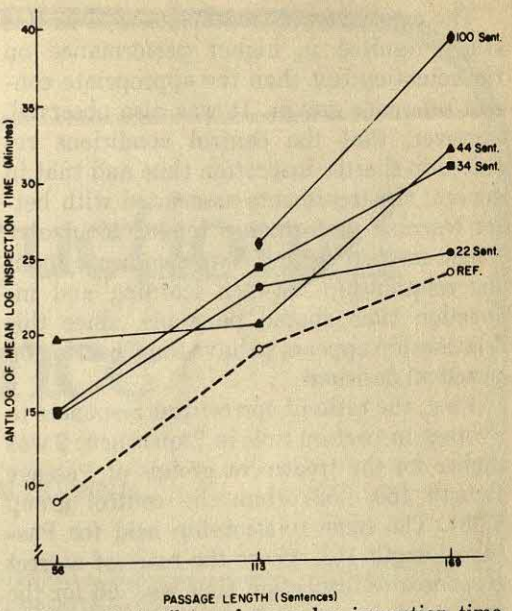


FIGURE 3. Antilogs of mean log inspection time for treatments involving various numbers of objective-relevant sentences as a function of passage length (Sent = sentences).

the following: (*a*) Providing instructional objectives increases the amount of objective-relevant material that is learned as well as the learning of incidental background material over the performance levels achieved by suitable control groups who were directed to learn as much as possible, and (*b*) specifically stated objectives resulted in better performance than more generally phrased objectives.

The finding that directions to learn particular content in a text also facilitates the acquisition of incidental material is at variance with results recently reported by Duchastel (1972). He found that subjects who were given objectives performed somewhat more poorly on incidental material than a nonobjective control group. The reason for the difference between the results of these experiments is not clear. Directions to attain specific goals may interfere with learning of incidental material when the number of objective-relevant sentences in a passage is relatively small. The density of goal-relevant sentences in the Duchastel study was only about 1%, whereas in our studies the densities ranged from 13% to 85%.

The experimental treatments used in this study resulted in higher performance on the retention test than the appropriate control reference groups. It was also observed, however, that the control conditions resulted in shorter inspection time and that in general the treatments associated with better learning performance tended to involve longer inspection time. Two comments about the relationship between learning and inspection time should be made, since this relationship appears to have some bearing on practical decisions.

First, the ratio of correct test responses to average inspection time in Experiment 2 was higher for the treatment groups of Passage Length 169 (.52) than the control group (.38). The same relationship held for Passage Length 113, where the ratio of correct responses to inspection time was .66 for the treatment groups and .44 for the control. This would suggest that the treatment is in both cases practically preferable over the control, since both learning score and the average rate of achievement of this learning score in time were greater for the treatments.

For Length 56, however, the ratio of correct intentional learning responses to inspection time is greater for the control (1.21) than the treatment (.76). Whether this implies that the control group is practically preferable over the treatment for Length 56 passages depends on the relationship between achievement and inspection time. If it is assumed that with increased opportunity for study, inspection time produces linear increases in test performance until mastery is achieved, then the control group appears preferable for the 56-sentence passage. This is because slope alone, that is, the ratio of correct response to inspection time, is sufficient to predict when mastery will occur. Assumptions of various curvilinear relationships between performance and inspection time may support other decisions. Assumptions of this type would not be unreasonable. Rothkopf (1968), for example, has found that for prose passages of approximately 750 and 1,500 words in length, the average number of correct test responses increased in a negatively accelerated manner with cumulative inspection time. If the relationship between performance and time increases monotonically but is negatively accelerated, the treatment with higher performance but lower performance/time ratio may be preferable to the other treatments if all treatments result in the same acquisition curves but vary in how much time is devoted to study when the opportunity occurs.

The two experiments reported here have not explored the effect on learned performance of the length of the list of objectives given to subjects as compared to the number of objective-relevant sentences in the passage. However, this is not a serious practical problem, since these two factors are in practice closely associated. They were also correlated in this study.

The present results, together with those previously reported by Rothkopf and Kaplan (1972), strongly suggest that when instructional objectives are known, substantial gains in instructional effectiveness may be obtained by communicating these objectives to students embarking on reading assignments.

REFERENCES

DUCHASTEL, P. C. *Incidental and relevant learning with instructional objectives.* (Technical Memo 66) Tallahassee: Flordia State University, Computer-Assisted Instruction Center, 1972.

ROTHKOPF, E. Z. Textual constraint as function of repeated inspection. *Journal of Educational Psychology*, 1968, **58**, 20–25.

ROTHKOPF, E. Z., & COKE, E. U. Learning about added sentence fragments following repeated inspection of written discourse. *Journal of Experimental Psychology*, 1968, **78**, 191–199.

ROTHKOPF, E. Z., & KAPLAN, R. An exploration of the effect of density and specificity of instructional objectives on learning from text. *Journal of Educational Psychology*, 1972, **63**, 295–302.

(Received August 22, 1973)

Instructional Science

AN INTERNATIONAL JOURNAL

Advisory Board

M. ADELSON
A. BJERSTEDT
A. CICOUREL
J. DONIO
H. FISCHER
N. L. GAGE
R. M. GAGNÉ
J. L. GEDYE
R. GLASER
D. JAMISON
A. KAUFMANN
H. KAY
L. de KLERK
F. KLIX
J. KONORSKI
P. MEREDITH
M. W. NEIL
G. PASK
A. RAPOPORT
R. J. SEIDEL
M. SMILANSKY
P. SUPPES
N. F. TALYZINA
S. TANAKA
M. TODA
A. TVERSKY
J. ZIELINSKI
K. ZINN

Under the editorship of:

GERALD F. BRIESKE, Kingston Polytechnic, Kingston-upon-Thames, Great Britain

BRIAN N. LEWIS and MICHAEL MacDONALD-ROSS, The Open University, Bletchley, Great Britain

ROULETTE WM. SMITH, The University of California, Santa Barbara, U.S.A.

INSTRUCTIONAL SCIENCE publishes fundamental research and review articles on all aspects of instructional science. The primary aim is to contribute to a deeper understanding of the nature, theory and practice of the instructional process and the learning to which it gives rise.

This quarterly journal cuts across the boundaries between the many disciplines in the behavioural, social and communication sciences. By doing so it facilitates communication between researchers and practitioners working on the same subject, but from different directions.

The 1974 subscription price (Volume 3) is Dfl. 88.00 or £11.80 or US$35.20

Prepaid personal subscriptions for private use are also available at Dfl. 45.00 or £6.00 or US$18.00.

Volume One (1972 and Two (1973) are still available at Dfl. 88.00.

For more details and sample copies, please write to

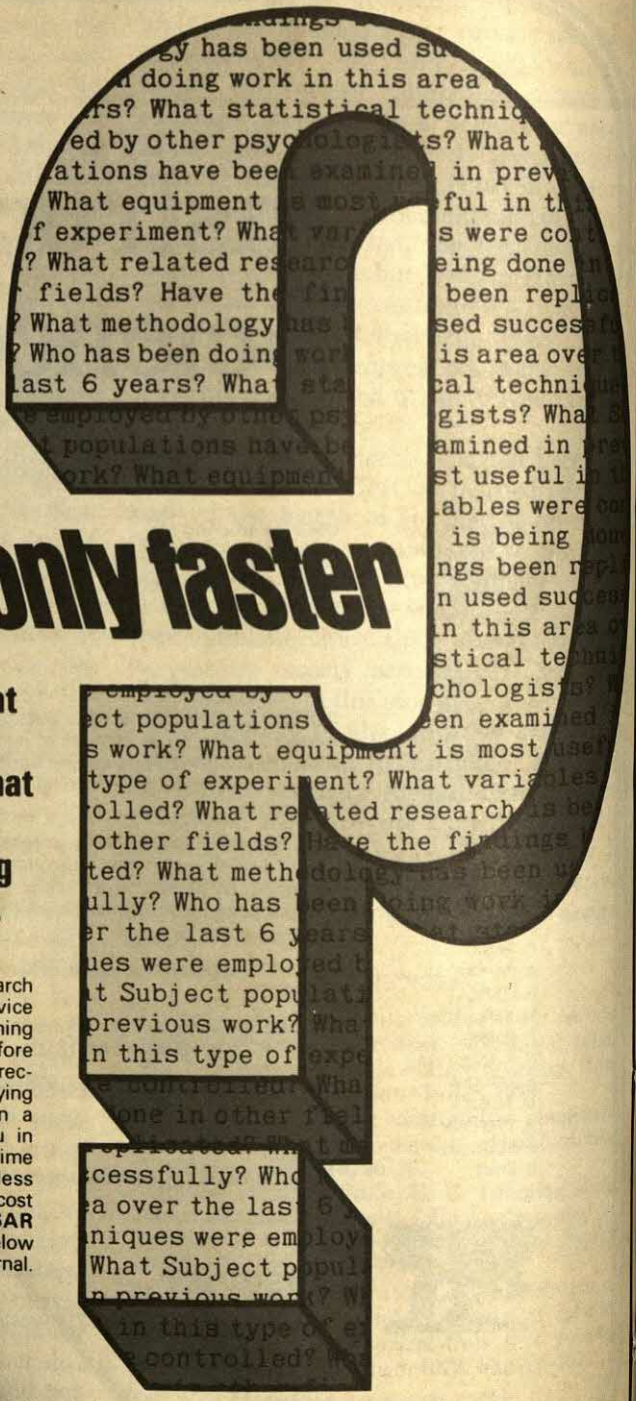


Elsevier

Journal Division, P.O. Box 211 Amsterdam - The Netherlands

5002E

If you can always find the psychological information you're looking for in the library stacks . . .



Pasar is only faster

but chances are great that Pasar will enable you to retrieve some material that would remain lost to the manual searcher devoting months to the same task.

PASAR (*Psychological Abstracts* Search and Retrieval) is an automated service that affords rapid bibliographic searching at a level of complexity never before possible. A computer printout of *PA* records (from 1967 to the present) satisfying the information need you specify on a **PASAR** Request Form is sent to you in about 2-3 weeks—a fraction of the time frequently required to perform a less satisfactory manual search. Average cost ranges $40-$60. To obtain a **PASAR** Request Form, write to the address below or consult a recent issue of the *PA* journal.

Psychological Abstracts Information Services
AMERICAN PSYCHOLOGICAL ASSOCIATION
1200 Seventeenth Street, N.W. Washington, D.C. 20036

A convenient guide to innovative curricula, training programs, model projects, and resources in elementary education.

A SOURCEBOOK OF ELEMENTARY CURRICULA PROGRAMS AND PROJECTS

Senior Editor **Samuel N. Henrie**

From the ALERT Information System of the Far West Laboratory for Educational Research and Development

Published by the **DOCENT CORPORATION,** Pleasantville, New York.

The national investment in educational research and development is yielding important new products that offer substantial improvements in teaching, learning, and educational management.

Unfortunately, most of the decision-makers in schools and colleges are either unaware of what is available or they lack sufficient information to make choices about them. For this reason many important new products are not yet in widespread use. This Sourcebook represents a major attempt to remedy this situation and hasten the flow of products of research and development into the stream of educational practice.

Each of the 300 reports in this oversized volume begins with a concise introduction providing general information. Then, using an outline format, information on subjects such as *Target Audience, Content Emphasis, Project Goals, Length of Use, Unit Sequencing, Instructional Method, Teacher's Role and Training, and Program Evaluation* are provided. Finally, a quick reference chart spells out the items required for the program: Quantities needed, Cost per item, and Replacement rate. Each report follows the same format so that a reader can easily compare the merits of different programs.

Directory to Alert Sourcebook

The Directory, a companion volume to the ALERT Sourcebook, provides a list of schools in which the Sourcebook programs are being used, and which are willing to receive visitors and/or answer inquiries.

ORDER FORM

Please enter my order for:

_____ copies of A SOURCEBOOK OF ELEMENTARY CURRICULA, PROGRAMS AND PROJECTS, and the DIRECTORY TO THE ALERT SOURCE BOOK at $12.40 for the two volumes

_____ copies of the SOURCEBOOK only, $9.95

_____ Bill me for the items ordered plus .90 to cover shipping

A check for _____ is enclosed (Docent will pay shipping charges)

Name _____

Address _____

City _____ State _____ Zip _____

DOCENT CORPORATION, 430 Manville Rd., Pleasantville, N.Y. 10570 [(914) 769-3629]

CHARLES C THOMAS · PUBLISHER

CLASSROOM DISCIPLINE: A Positive Approach *by* **Susan Bray Stainback,** *Hope Haven Children's Hospital, Jacksonville, and* **William Clarence Stainback,** *Univ. of North Florida, Jacksonville.* The primary focus of this book is on straightforward, specific, workable techniques for classroom control. The authors outline exactly what has to be done in the real, live classroom situation to control disruptive behavior. Viable alternatives to outmoded techniques of classroom control, such as threats and corporal punishment, are also provided. *'74, 180 pp., $5.50, paper*

VOCATIONAL REHABILITATION OF THE MENTALLY RETARDED: A Book of Readings. *Edited by* **Lloyd Keith Daniels.** *Central Connecticut State College, New Britain. (104 Contributors)* Included in this work are a wide variety of topics essential to broadening knowledge regarding the retarded. The articles were chosen for their logical and consistent arrangement of ideas and each selection either describes a program or presents a systematic procedure for resolving an issue or solving a problem of traditional concern to personnel engaged in the vocational rehabilitation of the mentally retarded. *'74, 648 pp. (7 x 10), 11 il., 60 tables, $22.50*

ORGANIZATION AND ADMINISTRATION OF EDUCATIONAL PROGRAMS FOR EXCEPTIONAL CHILDREN *by* **B. R. Gearheart,** *Univ. of Northern Colorado, Greeley.* Written for use as a basic college text for those training to become supervisors or administrators of special education programs, this volume should be of interest and value to all those who come in contact with exceptional children. The review of federal legislation for the handicapped and its relationship to the evolving role of federal leadership, special financing for programs, and the scope and function of both state and local special education administration add to the practical applicability of this volume. *'74, 272 pp., 1 table, $11.75*

TRAINING RETARDED BABIES AND PRESCHOOLERS *by* **Thomas F. Linde,** *V. A. Hospital, Knoxville, Iowa, and* **Thusnelda Kopp,** *Washoe County Association for Retarded Children, Reno, Nevada.* Based on the premise that most mentally retarded babies and preschoolers possess important potential to make productive contributions to their families, neighborhoods and communities, this volume is designed for public health personnel, psychologists, parents and others initiating and implementing activities for shaping basic perceptual, motor and social skills. *'74, 200 pp., 12 il., $12.75*

EFFECTIVE TEACHING METHODS FOR AUTISTIC CHILDREN *by* **Rosalind C. Oppenheim,** *The Rimland School for Autistic Children, Evanston, Illinois. Foreword by* **Bernard Rimland.** This reference book for parents, special educators and professionals in related disciplines emphasizes intruding on the child's withdrawal and overcoming resistance and negativism; establishing control, such as dealing with tantrums and disruptive behavior; the teacher-child relationship in the learning situation; and methods for teaching reading, writing and arithmetic. *'74, about 104 pp.*

THE RIGHT-TO-EDUCATION CHILD: A Curriculum for the Severely and Profoundly Mentally Retarded *by* **Donald G. Myers,** *Marywood College, Scranton, Pennsylvania, and* **Michael E. Sinco** *and* **Ellen Sommerton Stalma,** *both of Northeastern Intermediate Unit Nineteen, Scranton, Pennsylvania. Foreword by* **P. M. Mensky.** *(2 Contributors)* For teachers, college students, parents and those interested in working with the severely and profoundly mentally retarded, this book offers techniques based on well-established principles of learning. Step-by-step, carefully detailed procedures are provided for helping these children achieve better sensory development, motor development, self-care development and language readiness development. *'74, 248 pp., 41 il., $9.75*

Orders with remittance sent, on approval, postpaid

301-327 East Lawrence Avenue • Springfield • Illinois • 62717

Journal of Educational Psychology

VOLUME 66 NUMBER 4

AUGUST 1974

Duplicate 2 copies
to the Library
2/11/75

AMERICAN PSYCHOLOGICAL ASSOCIATION

EDITOR

JOANNA WILLIAMS, *Teachers College, Columbia University*

ASSOCIATE EDITOR

SAMUEL BALL, *Educational Testing Service*

ADVISORY EDITORS

PAUL R. AMMON, *University of California, Berkeley*
JERE BROPHY, *University of Texas at Austin*
C. VICTOR BUNDERSON, *Brigham Young University*
LEONARD S. CAHEN, *Educational Testing Service*
LOIS-ELLIN DATTA, *National Institute of Education*
FREDERICK B. DAVIS, *University of Pennsylvania*
FRANCIS J. DIVESTA, *Pennsylvania State University*
ROBERT EBEL, *Michigan State University*
DAVID ELKIND, *University of Rochester*
LAWRENCE T. FRASE, *Bell Laboratories*
N. L. GAGE, *Stanford University*
ROBERT GAGNÉ, *Florida State University*
ROBERT GLASER, *University of Pittsburgh*
MIRIAM L. GOLDBERG, *Teachers College, Columbia University*
EDMUND W. GORDON, *Teachers College, Columbia University*
VERNON C. HALL, *Syracuse University*
ARTHUR R. JENSEN, *University of California, Berkeley*
PHYLLIS A. KATZ, *City University of New York, Graduate School and University Center*
HERBERT J. KLAUSMEIER, *University of Wisconsin—Madison*
ROBERT E. LANA, *Temple University*
GERALD S. LESSER, *Harvard University*
JOEL R. LEVIN, *University of Wisconsin—Madison*
FREDERICK J. MCDONALD, *Educational Testing Service*
WILBERT J. MCKEACHIE, *University of Michigan*
FRANK B. MURRAY, *University of Delaware*
LAUREN B. RESNICK, *University of Pittsburgh*
WILLIAM D. ROHWER, JR., *University of California, Berkeley*
ROBERT ROSENTHAL, *Harvard University*
ERNST Z. ROTHKOPF, *Bell Laboratories*
S. JAY SAMUELS, *University of Minnesota*
SEYMOUR B. SARASON, *Yale University*
MELVYN I. SEMMEL, *Indiana University*
IRVING SIGEL, *Educational Testing Service*
HARRY SILBERMAN, *University of California, Los Angeles*
RICHARD E. SNOW, *Stanford University*
GEORGE G. THOMPSON, *Ohio State University*

JOURNAL STAFF

HAROLD P. VAN COTT, *Managing Editor*
ANITA DEVIVO, *Executive Editor and General Manager*
LINDA B. BAROODY, *Supervising Technical Editor*
ANN G. CORBETT, *Technical Editor*
LYNN D. SPIER, *Acting Manager, Publications Business Service*
ANNE REDMAN, *Subscription Manager*

Published bimonthly in one volume per year, six issues (February, April, June, August, October, and December), by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D. C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D. C., and additional mailing offices.

The **Journal of Educational Psychology** publishes original investigations and theoretical papers dealing with learning and cognition, especially as they relate to problems of instruction, and with the psychological development, relationships, and adjustment of the individual. Preference is given to studies of the more complex types of behavior, especially in or relating to educational settings. Journal articles pertain to all levels of education and to all age groups.

Manuscripts: Submit manuscripts in triplicate to the Editor, Joanna Williams, P.O. Box 51, Teachers College, Columbia University, New York, New York 10027, according to instructions provided elsewhere in the Journal.

Copyright and Permissions: All rights reserved. Written permission must be obtained from the American Psychological Association for copying or reprinting text of more than 500 words, tables, or figures. Permissions are normally granted contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of $10 per page of APA material. Permission fees are waived for authors who wish to reproduce their own articles, or when material is reproduced in limited quantities for instructional purposes. Address requests for permission to the Permissions Office, American Psychological Association, 1200 Seventeenth Street, N. W., Washington, D.C. 20036.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1974: $24 domestic, $25 foreign, $5 single issue. APA member rate: $10. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive galley proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Department, APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P. O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. Undelivered copies resulting from address changes will not be replaced; second-class mail is not forwarded unless subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

Volume 66 August 1974 Number 4

Another View of the Relation of Environment to Mental Abilities 457
CHESTER W. HARRIS AND DAVID L. McARTHUR

Another View of the Relation of Environment to Mental Abilities: A Reply 460
KEVIN MARJORIBANKS

Pseudo-Orthogonal and Other Analysis of Variance Designs Involving Individual-Differences Variables 464
LLOYD G. HUMPHREYS AND ALLEN FLEISHMAN

Reinforcement Practices of Black and White Teachers in Integrated Classrooms. 473
ROBERT BYALICK AND DONALD N. BERSOFF

Incidental and Relevant Learning with Instructional Objectives 481
PHILIPPE C. DUCHASTEL AND BOBBY R. BROWN

Effects of Inserted Questions on Retroactive Inhibition in Meaningful Verbal Learning 486
BASIL S. WALKER

The Algorithmic Approach to Curriculum Construction: A Field Test in Mathematics 491
WALTER EHRENPREIS AND JOSEPH M. SCANDURA

Internalization of Filmic Schematic Operations in Interaction with Learners' Aptitudes 499
GAVRIEL SALOMON

Effects of Nonparticipant Observation on Teacher and Pupil Classroom Behavior 512
MICHAEL MERCATORIS AND W. EDWARD CRAIGHEAD

Origins of Teachers' Attitudes toward Young Children 520
SHERRY WILLIS AND JERE BROPHY

Differential Effects of Reward versus No-Reward Instructions on the Creative Thinking of Two Economic Levels of Elementary School Children 530
ROGER A. JOHNSON

Effects of the Availability of Objectives and/or Rules on the Learning Process .. 534
PAUL F. MERRILL

Influence of Mode of Presentation, Ethnicity, and Social Class on Teachers' Evaluations of Students 540
MARY JENSEN AND LAWRENCE B. ROSENFELD

Initial Investigations into Academic Survival-Skill Training, Direct Instruction, and First-Grade Achievement 548
HYMAN HOPS AND JOSEPH A. COBB

Signal Systems of Lesson Settings and the Task-Related Behavior of Preschool Children 554
JACOB S. KOUNIN AND PAUL V. GUMP

Intelligence and Transfer: Aptitude by Treatment Interactions 563
GRAHAM R. SKANES, ARTHUR M. SULLIVAN, EDWARD J. ROWE, AND ELIZABETH SHANNON

(Continued on next page)

Diary N. 848
27 | 1 | 75

Cognitive Style and Reasoning about Speed..... 569
LINNEA C. EHRI AND IRENE M. MUZIO

"Environments" of British Commonwealth Universities..... 572
JAMES M. RICHARDS, JR.

Unfamiliar Stimulus Terms in Children's Paired-Associate Learning..... 580
ROBERT E. DAVIDSON, SHERRY A. B. PERRY, AND PHOEBE K. BAKER

Validity of Student Evaluation of Teaching and the Characteristics of Successful Instructors..... 584
ARTHUR M. SULLIVAN AND GRAHAM R. SKANES

Multidimensional Scaling of Concept Learning in an Introductory Course.... 591
HOWARD WAINER AND KENNETH KAYE

Some Problems in the Application of Achievement Motivation to Education: The Assessment of Motive to Succeed and Probability of Success..... 599
CHARLES B. SCHULTZ AND MICHAEL POMERANTZ

Course Evaluation: When?..... 609
NEIL A. CARRIER, GEORGE S. HOWARD, AND WILLIAM G. MILLER

Effects of Instructional Objectives Used as Orienting Stimuli or as Summary/Review upon Prose Learning..... 614
ROBERT KAPLAN AND FRANCINE G. SIMMONS

Training for Thematic-Fantasy Play in Culturally Disadvantaged Children: Preliminary Results..... 623
ELI SALTZ AND JAMES JOHNSON

Announcement..... 553

List of Manuscripts Accepted..... 622

SUGGESTIONS TO CONTRIBUTORS

In preparing articles for publication in the Journal, authors are strongly advised to follow the general directions given in the *Publication Manual of the American Psychological Association* (1967 Revision). All articles must be preceded by an abstract of 100–120 words. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833). Special attention should also be given to the section on the preparation of the references (*Publication Manual*, p. 39). In reference lists, give journal titles in full; do not abbreviate. *All copy must be double spaced, including the references.* All manuscripts should be submitted in *triplicate*, one of which should be an original typed copy. Since the reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the identity of the author or his affiliation should be on separate pages. Every effort should be made by the author to see that the manuscript itself contains no clues as to his identity. Dittoed and mimeographed copies are not acceptable and will not be considered. Original figures should be high-contrast, glossy, black and white prints of professional quality. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail and to check carefully the typing of the final copy.

Journal of Educational Psychology
1974, Vol. 66, No. 4, 457-459

ANOTHER VIEW OF THE RELATION OF ENVIRONMENT TO MENTAL ABILITIES

CHESTER W. HARRIS[1] AND DAVID L. McARTHUR

University of California, Santa Barbara

Correlations between environment measures and mental ability test performance measures presented by Marjoribanks were examined for the presence of a factor, or factors, common between the two sets of measures. A single latent common factor was found to account, by itself, for all but a small portion of the total variance. This finding supports the conclusion that the environment measures are not differentially related to test performance but rather differ only in their correlation with the single latent factor.

The purpose of this note is to propose another interpretation of the data presented by Marjoribanks (1972) relating environment measures (press and status) to mental ability test performance. Marjoribanks used his data to search for "subenvironments or subsets of environmental forces" that are related to each of four abilities as measured by the SRA Primary Mental Abilities subtests. He presents correlations between eight "environmental forces" (press for achievement, press for activeness, etc.) and the mental ability test scores, and he also gives correlations between "global" or status indicators of the environment (education of each parent, occupation of father, etc.) and these same mental ability test scores. He found, in general, higher correlations between press variables and test scores than between status variables and test scores. He also found that the spatial and reasoning subtests correlated less highly with the press variables than did the verbal and number subtests.

We wish to answer a slightly different question, using Marjoribanks' data.[2] The question is this: How many latent variables are required to "account for" or reproduce the correlations between environment measures (press and status) and mental ability test scores? It can be seen in advance that the answer to this question may be critical to the interpretation of the observed differences in level of correlation. For example, it may be possible that the two types of environment measures differ in their correlation with the test scores primarily because the two types of measures (press and status) differ in the variance of their true score components and not because the true score components of the two types of environment measures are measuring different constructs or latent measures. The analysis by Marjoribanks does not permit one to answer this kind of question.

In order to answer the question of how many latent variables are required to account for the cross-correlations between environment measures and test scores, we used Tucker's interbattery factor analysis model (Tucker, 1958). Tables 3 and 5 of the original article (Marjoribanks, 1972) are the relevant ones for this analysis, but recall that some correction of the data presented originally is required. One can simply set Table 5 below Table 3 to give the complete 14×4 matrix of cross-correlations between environment and test scores. (We eliminated the "Total" column, since the total score is simply a linear composite of the four subscores: verbal, number, spatial, and reasoning.) This consolidated matrix of cross-

[1] Requests for reprints should be sent to Chester W. Harris, Department of Education, University of California, Santa Barbara, California 93106.

[2] We are indebted to Kevin Marjoribanks for supplying us with corrected correlation matrices after we found a slight error in the published data.

TABLE 1

CORRELATIONS OF VARIABLES WITH INTERBATTERY FACTOR

Variable	Correlation
Verbal ability	.86
Number ability	.86
Spatial ability	.59
Reasoning ability	.66
Press for achievement	.70
Press for activeness	.49
Press for intellectuality	.60
Press for independence	.39
Press for English	.43
Press for "ethlanguage"	.31
Father dominance	.15
Mother dominance	.17
Education of father	.33
Education of mother	.17
Occupation of father	.43
Number of children	−.29
Crowding ratio	−.32
Ordinal position	−.24

correlations then can be analyzed in the manner described by Tucker to give the correlations of each of the variables with one or more factors common to the two batteries (interbattery factors).

For this particular set of data, the question of how many interbattery factors are required was answered by examining the associated eigenvalues (generalized variances), and it was found that the first eigenvalue (5.093) closely approximated the total amount of generalized variance (5.241), thus indicating that only one interbattery factor should be postulated. The remaining three nonzero eigenvalues were all less than .110. Our first finding, then, is that only a single latent variable is common to the environment and the test variables. This is an important conclusion, since it says that within the context of their association with the four test variables, the two types of environment measures differ primarily in the magnitude of their correlation with a single latent variable but do not measure different aspects of what is common to both the environment and the test variables. A contribution that Marjoribanks has made, therefore, is that of designing "press" variables that are factorially similar to the "status" variables but are better predictors of mental test scores. So far as these measures, used with this sample, are concerned, the 14 environment measures all tap the same aspect of what is common with the test scores. Similarly, we conclude that there is no evidence of "subsets of environmental forces" that relate differentially to the four test score variables. In their relation with these environment measures, the four test score variables differ only in the magnitude of their correlation with the single latent variable that is common to the two batteries.

The correlation of each variable in each of the two batteries with the single interbattery factor can also be calculated. These are given in Table 1, and they indicate the differences referred to above. From these correlations, it is clear that the single environment variable that is correlated most highly with this interbattery factor is "press for achievement." Its correlation with each of the four test score variables can be reproduced approximately by multiplying its coefficient in Table 1 with the coefficient for each of the four test score variables. Thus, Marjoribanks (1972, Table 3) gives the correlation of press for achievement with Verbal as .66, and our values give .86 times .70 equals .6020 as the approximation (see Table 1).

It should be emphasized that this single interbattery factor provides a very good reproduction of the cross-correlations between environment measures and test scores. However, it does not necessarily reproduce the within-battery correlations.[3] This is simply to say that, for example, the environment measures have further factors in common with each other beyond (and independent of) the single factor in common with the test score variables. This is quite reasonable, since measures like number of children in the family and crowding ratio are measuring much the same thing and thus tend to have a factor in common independent of their relationship with test score variables. We are stressing that the interbattery analysis focuses attention on the factors held in common by the two batteries (environment

[3] These within-battery correlations were also kindly supplied by Kevin Majoribanks.

and test scores) and must be supplemented by further analyses if all the factors characteristic of both batteries are desired. Our interest here was in this narrow focus, since it tends to support conclusions that differ somewhat from those given in the original study.

REFERENCES

MARJORIBANKS, K. Environment, social class, and mental abilities. *Journal of Educational Psychology*, 1972, **63**, 103–109.

TUCKER, L. R. An interbattery method of factor analysis. *Psychometrika*, 1958, **23**, 111–136.

(Received January 15, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 460-463

ANOTHER VIEW OF THE RELATION OF ENVIRONMENT TO MENTAL ABILITIES:

A REPLY

KEVIN MARJORIBANKS[1]
University of Oxford, England

Harris and McArthur (1974) concluded that Marjoribanks' (1972) measure of the learning environment of the home was not differentially related to mental ability test performance. It is argued that the original study did not propose that the two types of environmental measures were assessing different constructs and that Harris and McArthur's interbattery factor analysis of Marjoribanks' data actually supports the original proposition. It is suggested that Walberg and Marjoribanks' (1973) canonical correlation analysis of the data, which revealed a two-dimensional space that is common between the abilities and the environmental measures, provides a better explanation of the data than Harris and McArthur's analysis.

Harris and McArthur have added little to our understanding of the relation between environments and mental abilities by their extra analysis of the data which appeared in Marjoribanks (1972), but they have presented an interesting statistical discussion.

The purpose of the original study was to examine the relationship between mental ability scores and different indicators of the same construct, namely, home environment. It was not proposed that the two types of environment measures were assessing different constructs. However, in relation to the global classificatory environmental characteristics, it is proposed that the psychosocial parent–child interaction variables (*a*) have greater diagnostic and functional value for the educator and (*b*) facilitate our understanding of the basic nature and function of the mental abilities themselves and of the characteristics of the environmental conditions which influence mental ability development. Therefore, by indicating that the environmental press variables are factorially similar to the social status variables and are more highly correlated with the interbattery factor, the new analysis provides support for the propositions of the original study.

In relation to the comment on the association between the environment measures and differential mental abilities, a potentially more useful statistical analysis has been completed in collaboration with Professor Walberg of the University of Illinois at Chicago Circle. As the interbattery factor analysis conducted by Harris and McArthur (1974) indicates, the results published in the original study support only one dimension in common between mental abilities and the environmental indicators. However, a canonical correlation analysis of the complete set of intercorrelations between mental abilities, social status indicators, and the environmental force scores (see Table 1), indicates a two-dimensional space that is common between the abilities and the environment measures.

When canonical correlations are computed between the eight environmental forces and the six global environmental indicators as a set of predictor variables,

[1] Requests for reprints should be sent to Kevin Marjoribanks, Department of Educational Studies, Oxford University, 15 Norham Gardens, Oxford OX 2 6 py England.

TABLE 1

INTERCORRELATIONS BETWEEN MENTAL ABILITIES, GLOBAL ENVIRONMENTAL INDICATORS, AND ENVIRONMENT FORCE SCORES

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Verbal ability	—																	
2. Number ability	.66	—																
3. Reasoning ability	.39	.44	—															
4. Spatial ability	.34	.38	.43	—														
5. Father occupation	.43	.30	.29	.31	—													
6. Father education	.29	.27	.22	.26	.79	—												
7. Mother education	.38	.32	.15	.20	.60	.63	—											
8. Number children	−.32	−.33	−.03	−.04	.13	−.09	−.25	—										
9. Ordinal position	−.27	−.26	−.04	−.04	.12	−.11	−.26	.87	—									
10. Crowding ratio	−.34	−.34	−.09	−.07	.23	−.17	−.34	.83	.73	—								
11. Press for achievement	.66	.66	.39	.28	.56	.50	.53	−.46	−.40	−.51	—							
12. Press for activeness	.52	.41	.26	.22	.45	.48	.49	−.30	−.26	−.33	.72	—						
13. Press for intellectuality	.61	.53	.31	.26	.51	.50	.54	−.35	−.31	−.40	.80	.80	—					
14. Press for independence	.42	.34	.22	.10	.36	.37	.32	−.06	−.08	−.07	.40	.46	.48	—				
15. Press for English	.50	.27	.28	.18	.34	.32	.43	−.16	−.15	−.15	.52	.63	.61	.49	—			
16. Father dominance	.16	.10	.11	.09	.20	.26	.30	−.20	−.22	−.22	.32	.43	.38	.11	.23	—		
17. Mother dominance	.21	.15	.04	.04	.20	.23	.21	−.06	.04	.01	.26	.30	.37	.35	.32	−.48	—	
18. Press for ethlanguage	.35	.24	.18	.09	.36	.44	.46	−.22	−.22	−.31	.51	.64	.62	.41	.45	.34	.21	—

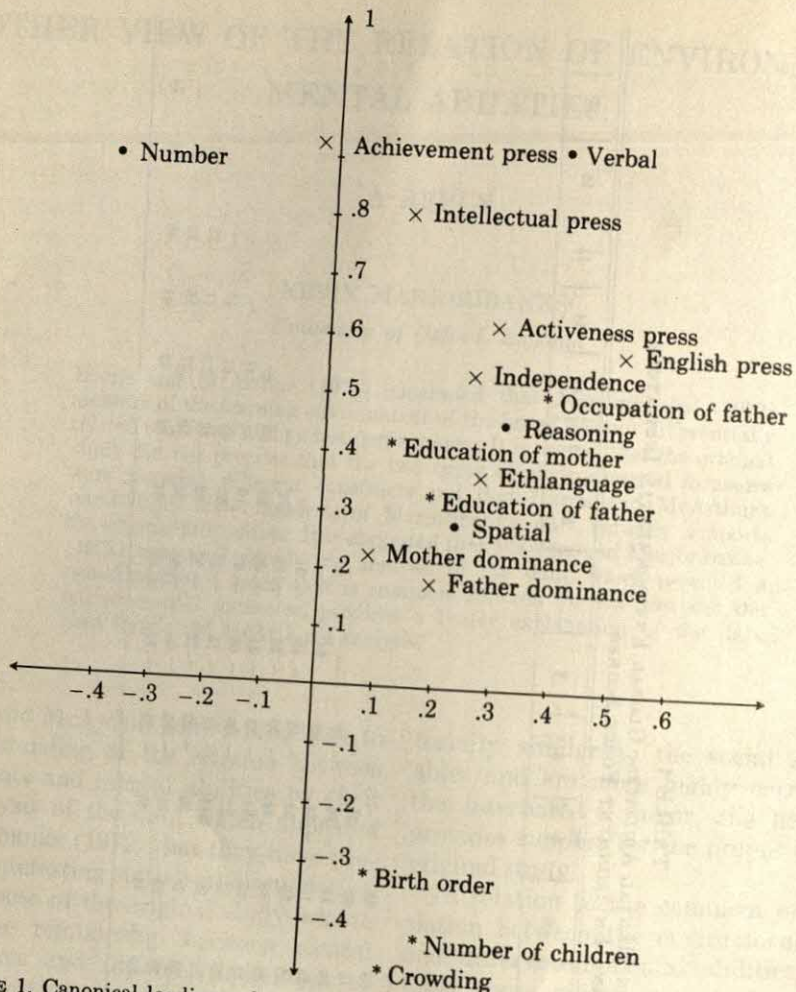


FIGURE 1. Canonical loadings of environmental and mental abilities measures. (Legend: × = environmental press, * = global environment indicator, ● = mental ability measure.)

and the four mental abilities as a set of criterion variables, significant canonical correlations of .781 and .462 are obtained (probabilities less than .001 and .005, respectively). The standard errors of these correlations are .029 and .058, respectively.

The relations between the environment measures, the mental abilities, and the two canonical variates are shown in Figure 1. The values used in the figure, canonical loadings, are the correlations between the original measures and canonical variates of the same battery: predictor measures with predictor canonical variates and criterion measures with criterion canonical variates (Bock & Haggard, 1968; Darlington, Weinberg, & Walberg, 1973).

The canonical loadings indicate that with respect to the first canonical variate, verbal and number abilities and, to a lesser extent, reasoning ability, are more closely associated with the environmental force scores than is spatial ability. The high loadings on the first canonical variate also indicate that the environmental forces contribute more strongly to the prediction of the abilities than do the social-status indicators and the family structure characteristics.

After removing the variance of the first canonical variate from predictors and criteria, the loadings on the second variate reveal that the social status indicators and the environmental forces are significantly related to differentially developed abilities (Walberg & Marjoribanks, 1973). It should be noted that the second variate explains

TABLE 2

CORRELATIONS BETWEEN THE TWO CANONICAL PREDICTOR VARIATES AND FOUR MENTAL ABILITIES

Ability	Simple r		Corrected R	% total variance
	Variate 1	Variate		
Verbal	.72***	.15**	.73***	52.8
Number	.70***	−.18***	.73***	52.8
Spatial	.28***	.10	.29***	8.1
Reasoning	.40***	.14*	.42***	17.5

* $p < .05$.
** $p < .01$.
*** $p < .001$.

less variance in the mental ability scores than does the first variate (see Table 2). High ratings on press for English, father occupation, press for ethlanguage, and to a lesser extent, press for activeness and father dominance, are associated on the second variate with high scores on verbal, reasoning, and spatial abilities, but associated with lower number ability scores. The results of the canonical analysis suggest the hypothesis that environmental forces may operate selectively to develop certain potential abilities and to leave others relatively underdeveloped.

By involving all the data from the original study, a canonical correlation analysis reveals a differential relation to the predictor variables between number ability and the other three abilities, thus providing a different explanation of the data than that provided by an interbattery factor analysis.

REFERENCES

BOCK, R. D., & HAGGARD, E. A. The use of multivariate analysis of variance in behavioral research. In. D. K. Whitla (Ed.), *Handbook of measurement and assessment in behavioral sciences*. Reading, Mass.: Addison-Wesley, 1968.

DARLINGTON, R., WEINBERG, S., & WALBERG, H. J. Canonical variate analysis and related techniques. *Review of Educational Research*, 1973, **43**, 433–454.

HARRIS, C. W., & MCARTHUR, D. L. Another view at the relation of environment to mental abilities. *Journal of Educational Psychology*, 1974, **66**, 457–459.

MARJORIBANKS, K. Environment, social class, and mental abilities. *Journal of Educational Psychology*, 1972, **63**, 103–109.

WALBERG, H. J., & MARJORIBANKS, K. Differential mental abilities and home environment: A canonical analysis. *Developmental Psychology*, 1973, **9**, 363–368.

(Received March 6, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 464–472

PSEUDO-ORTHOGONAL AND OTHER ANALYSIS OF VARIANCE DESIGNS INVOLVING INDIVIDUAL-DIFFERENCES VARIABLES

LLOYD G. HUMPHREYS[1] AND ALLEN FLEISHMAN

University of Illinois at Urbana-Champaign

In a critique of a paper by Humphreys and Dachler, Fischbach and Walberg advised research workers who use individual-differences variables in orthogonal analysis of variance designs to obtain equal *Ns* in each cell and not to worry about population *Ns*. This is thoroughly misleading advice and is based upon an inadequate model of components of variance in individual-differences measures. On the basis of present analyses of the problem, it is concluded that the analysis of variance is an awkward, inefficient statistical model in these cases and that correlational analysis has many advantages for such problems. Some of the literature involving the pseudo-orthogonal design advocated by Fischbach and Walberg can be salvaged, when properly interpreted, but other research involving this method should be discarded and a fresh start should be made with adequate design and methods of analysis.

In their discussion of Jensen's (1968) theory of intelligence, Humphreys and Dachler (1969a, 1969b) characterized the use by Jensen, and many others, of two or more dichotomized, individual-differences variables in an analysis of variance design with equal *Ns* in each of the cells as a pseudo-orthogonal design. Humphreys and Dachler claimed that whenever such variables are correlated (which is quite literally always) in the population, the use of the pseudo-orthogonal design biases the differences in means for the main effects relative to the differences in those means that would be obtained in a single-factor experiment. For example, if measures of intelligence and socioeconomic status are each dichotomized and are used as the independent variables in an analysis of variance of rote-learning scores, the contributions to variance of the independent variables are changed by the use of equal *Ns* in the direction expected for partial correlations. Thus, the contribution to variance of socioeconomic status in the two-factor analysis is substantially smaller, as a function of the pattern of intercorrelations, than in the single-factor analysis. Jensen, and others, have not recognized this fact.

Equal versus Population Ns

Humphreys and Dachler (1969a, 1969b) contrasted the estimation of main effects in which means were equally weighted (the pseudo-orthogonal analysis) with an analysis which utilized the weighting of means by estimates of the *Ns* in the population. Unfortunately, however, they committed two errors, one major and one minor, in their discussion. The minor error was their failure to take into account the bias in either weighting procedure when a variable is dichotomized by the selection of extreme cases only, for example, upper and lower quartiles. The major error was their failure to give sufficient emphasis to their recommendation that investigators use correlational analysis rather than analysis of variance for individual-differences variables. They contrasted two methods of estimating main ef-

[1] Requests for reprints should be sent to Lloyd G. Humphreys, Department of Psychology, University of Illinois, Champaign, Illinois 61820.

fects in order to demonstrate how conclusions differ and, in doing so, almost lost sight of the main thrust of their argument, which was to use correlational analysis and to avoid the analysis of variance for such variables.

Fischbach and Walberg Criticisms

Subsequently, Fischbach and Walberg (1971) criticized the recommendations of Humphreys and Dachler (1969a, 1969b). They concluded that the procedure of weighting means by estimates of the *N*s in the population was biased and that the pseudo-orthogonal design with equal *N*s in the four cells produced unbiased estimates. These conclusions are clearly erroneous. While the Humphreys and Dachler estimates were biased, because the middle 50% of the cases had been omitted from the distributions of both independent variables, that bias is unrelated to the weighting-by-*N* procedure as such.

Because Fischbach and Walberg (1971) were incorrect in their analysis of the two designs, they also concluded by giving investigators thoroughly undesirable advice. Their concluding paragraph summarized their position:

> Moreover, this model also implies that when designing a study there is no need to be concerned about the relative numbers of persons in each of the four populations. The best and most economical method of gaining information about the parameters is to sample each group equally (unless, of course, there are sampling cost differentials). Thus, the question of what the joint distribution of IQ and socioeconomic status is like becomes an irrelevant issue [p. 80].

Fischbach and Walberg Errors

Fischbach and Walberg (1971) made two important errors in their development: one of omission and one of commission. With respect to the former, they failed to note that the arbitrary equalization of *N*s for individual-differences variables in order to impose orthogonality on the independent variables represents only an approximation to any of the three least squares solutions described by Overall and Spiegel (1969) for the problem of unequal *N*s. It is not obvious that equalizing *N*s experimentally is the equivalent of allowing for unequal *N*s statistically when individual-differences data are involved. This problem, it should be noted, did not occur to the latter authors, since they were concerned with experimental data. Unequal *N*s in experimental data carry no psychological meaning.

The difference between experimental data involving unequal *N*s and individual-differences data lies at the heart of the error of commission of Fischbach and Walberg. An orthogonal components of variance model is simply not an appropriate model for individual-differences variables. Instead, a model like the multiple-factor model or multiple regression model is appropriate. The orthogonal components of variance model does not include a socioeconomic status component in the variance of a dependent variable such as rote learning in a single-factor design in which intelligence is the single factor. Yet, it is practically impossible to reduce the socioeconomic status component present in the rote-learning scores to zero. Furthermore, an approximate reduction to zero variance of true scores, obtained by holding constant the obtained score on socioeconomic status, distorts relationships of other individual-differences measures with the rote-learning scores. Independence cannot be attained experimentally for individual-differences variables.

Theoretical Discussion

Basic Argument

The initial background for the above assertions can be obtained from a very elementary line of reasoning. If two variables are not orthogonal and each is dichotomized at the median, the *N*s in the four quadrants will necessarily be unequal. If one wishes to obtain the mean of a row or column, this is the equivalent of obtaining the mean of two samples combined. The formula is as follows:

$$\frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

As long as the separate means are unequal,

the combined mean is equal to $\bar{X}_1 + \bar{X}_2/2$ if, and only if, $N_1 = N_2$. Therefore, the two procedures will produce different estimates of main effects for individual-differences variables. The size of the difference in estimate depends, of course, on the size of the correlation between the variables and the points of dichotomization. Investigators *must* worry about which procedure they use. Furthermore, when one of the two procedures is clearly correct by a simple mathematical test as above and the other is incorrect by that test, a more complex mathematical development that seemingly reverses this conclusion should be viewed with caution. It is not surprising to find under these circumstances that the more complex mathematics depend on an assumption that leads to a special, limited case.

Interrelationships of Various Methods

A theoretical analysis of various ways of handling the problem of unequal *N*s by means of the analysis of variance is now required. Involved are the three methods of Overall and Spiegel (1969) and the weighting-by-*N* procedure of Humphreys and Dachler (1969a, 1969b). While the use of equal *N*s in the pseudo-orthogonal design can be considered a fifth procedure, its precise relationships with the other four have not been worked out analytically. Under favorable circumstances with respect to choice of dichotomization, which will be discussed later, it approximates one of the least squares methods.

The four analysis of variance methods are also related to traditional correlational analysis. To show this, it is helpful to use two correlational statistics, relating independent to dependent variables, derived from the analysis of variance. In order to equate these analysis of variance statistics with the usual correlational analysis, however, sample statistics rather than population estimates are presented in both instances. For a population estimate in the data which follow, the mean square for error should be subtracted from the numerator and added to the denominator for both omega-squared (Hays, 1963) and epsilon-squared (Peters & Van Voorhis, 1940).

$$\omega^2 = \frac{\text{sum of squares particular source}}{\text{sum of squares total}}$$

$$\epsilon^2 = \frac{\text{sum of squares particular source}}{\text{sum of squares particular source} + \text{sum of squares error}}$$

For convenience and ease of understanding, the discussion will assume only two levels each of two independent variables and the omission of no middle cases from the analysis. The relationships can easily be generalized to more than two independent variables and to more than two levels, but they cannot as readily be generalized to exclusion of middle cases. The above measures of the degree of relationship in the sample derived from the analysis of variance will be compared with squared partial and part correlations computed from point biserials between independent and dependent variables and phi coefficients between the independent variables.

Table 1 presents the interrelationships among the four analysis of variance approaches and the relationship of each to correlational analysis. The first three methods are identical with those described by Overall and Spiegel (1969). The fourth method is the weighting-by-*N* procedure which is presented in the symbols used by Overall and Spiegel. Definitions of various "error" or residual terms required by the relationships with correlational analysis are given in Footnote *a*, Table 1.

Since the reader can readily supply the equivalents of omega-squared and epsilon-squared written in terms of the analysis of variance components specified in the formulae presented, only the equivalent, squared point-biserial correlations and their derivatives appear in the appropriate columns. Setting off certain symbols by means of parentheses indicates part correlations, while the same expressions without the parentheses indicate partial correlations. The symbol $\underline{AB}$ indicates the interaction between the independent variables, while X denotes the dependent variable.

The omega-squared statistic requires the total sum of squares in the denominator in all cases. Thus, except when estimates of population values are involved, there is no

TABLE 1

SUMMARY OF RELATIONSHIPS AMONG FOUR LEAST SQUARES ANALYSIS OF VARIANCE METHODS FOR USE WITH UNEQUAL *Ns* AND COMPARISON OF EACH WITH CORRELATIONAL ANALYSIS

Source	*df*	*SS*	ω^2	Error[a]	ϵ^2	Error[a]	ϵ^2
Method 1							
A	1	$SS_T[R^2(\alpha_i,\beta_j,\alpha\beta_{ij}) - R^2(\beta_j,\alpha\beta_{ij})]$	$r^2_{X(A\cdot B\underline{AB})}$	1	$r^2_{XA\cdot B\underline{AB}}$		
B	1	$SS_T[R^2(\alpha_i,\beta_j,\alpha\beta_{ij}) - R^2(\alpha_i,\alpha\beta_{ij})]$	$r^2_{X(B\cdot A\underline{AB})}$	1	$r^2_{XB\cdot A\underline{AB}}$		
AB	1	$SS_T[R^2(\alpha_i,\beta_j,\alpha\beta_{ij}) - R^2(\alpha_i,\beta_j)]$	$r^2_{X(\underline{AB}\cdot AB)}$	1	$r^2_{X\underline{AB}\cdot AB}$		
Total	$n-1$	SS_T					
Method 2							
A	1	$SS_T[R^2(\alpha_i,\beta_j) - R^2(\beta_j)]$	$r^2_{X(A\cdot B)}$	2(a)	$r^2_{XA\cdot B}$	1	$r^2_{XA\cdot B}[(1 - r^2_{X\cdot B})/(1 - r^2_{X\cdot B\underline{AB}})]$
B	1	$SS_T[R^2(\alpha_i,\beta_j) - R^2(\alpha_i)]$	$r^2_{X(B\cdot A)}$	2(a)	$r^2_{XB\cdot A}$	1	$r^2_{XB\cdot A}[(1 - r^2_{X\cdot A})/(1 - r^2_{X\cdot A\underline{AB}})]$
AB	1	$SS_T[R^2(\alpha_i,\beta_j,\alpha\beta_{ij}) - R^2(\alpha_i,\beta_j)]$	$r^2_{X(\underline{AB}\cdot AB)}$	1	$r^2_{X\underline{AB}\cdot AB}$		
Total	$n-1$	SS_T					
Method 3							
A	1	$SS_T[R^2(\alpha_i)]$	r^2_{XA}	3(a)	r^2_{XA}	1	$r^2_{XA}[1/(1 - r^2_{X\cdot B\underline{AB}})]$
B	1	$SS_T[R^2(\alpha_i,\beta_j) - R^2(\alpha_i)]$	$r^2_{X(B\cdot A)}$	2(a)	$r^2_{XB\cdot A}$	1	$r^2_{XB\cdot A}[(1 - r^2_{X\cdot A})/(1 - r^2_{X\cdot A\underline{AB}})]$
AB	1	$SS_T[R^2(\alpha_i,\beta_j,\alpha\beta_{ij}) - R^2(\alpha_i,\beta_j)]$	$r^2_{X(\underline{AB}\cdot AB)}$	1	$r^2_{X\underline{AB}\cdot AB}$		
Total	$n-1$	SS_T					
Method 4							
A	1	$SS_T[R^2(\alpha_i)]$	r^2_{XA}	3(a)	r^2_{XA}	1	$r^2_{XA}[1/(1 - r_{X\cdot B\underline{AB}})]$
B	1	$SS_T[R^2(\beta_j)]$	r^2_{XB}	3(b)	r^2_{XB}	1	$r^2_{XB}[1/(1 - r^2_{X\cdot A\underline{AB}})]$
AB	1	$SS_T[R^2(\alpha\beta_{ij})]$	$r^2_{X\underline{AB}}$	3(c)	$r^2_{X\cdot\underline{AB}}$	1	$r^2_{X\cdot AB}[1/(1 - r^2_{X\cdot AB})]$
Total	$n-1$	SS_T					

Note. Setting off certain symbols by means of parentheses indicates part correlations, while the same expressions without the parentheses indicate partial correlations. The symbol $\underline{AB}$ indicates the interaction between the independent variables, while *X* denotes the dependent variable.

[a] Numbers in this column refer to the following error or residual terms:

Error$_1$ ($df = n - 4$): $SS_T[1 - R^2(\alpha_i,\beta_j,\alpha\beta_{ij})]$;

Error$_2$ ($df = n - 3$): $^{(a)}1 - R^2(\alpha_i,\beta_j)$, $^{(b)}1 - R^2(\alpha_i,\alpha\beta_{ij})$, $^{(c)}1 - R^2(\beta_j,\alpha\beta_{ij})$;

Error$_3$ ($df = n - 2$): $^{(a)}1 - R^2(\alpha_i)$, $^{(b)}1 - R^2(\beta_j)$, $^{(c)}1 - R^2(\alpha\beta_{ij})$.

problem with respect to choice of error term. A choice of error term is required, however, for epsilon-squared. The error used by Overall and Spiegel (1969) does not lead to ordinary correlational statistics in all cases. Other error terms must be substituted so that the correlations will be meaningful ones for individual-differences variables. These are reported in the right-hand epsilon-squared column. When unequal *N*s carry no psychological meaning, however, as is true for experimentally controlled variables, the "odd-ball" correlations in the righthand column which use the Overall and Spiegel estimate of error are meaningful. For experimental data, the operations on the standard deviations of the dependent variable called for by the equations are carried out very simply, that is, by omission of the other independent variables from a subsequent experiment.

For individual-differences variables, the operations specified by the equations utilizing the minimum error term *cannot* be made operational. The results are analogous to a correction for attenuation, which places the standard deviation of true scores in the denominator of the correlation in place of the standard deviation of obtained scores. Measurement error variance can be more readily manipulated, however, than the variance in a dependent variable associated with individual-differences variables other than the one under study.

The choice of error term is also important in hypothesis testing. When working with individual-differences variables, the same error terms used in computing the correlations must be used in tests of significance. When variables are under experimental control, on the other hand, the unequal *N*s carry no psychological meaning, and the minimum error term of Overall and Spiegel (1969) is appropriate.

Advantages of Correlational Analysis

The information in Table 1 illustrates very well the most important reason for the use of correlational analysis when dealing with individual-differences variables, namely, the ease and flexibility of correlational analysis. Four different methods for the analysis of data involving unequal *N*s produce different results, but these different results can all be obtained quickly and easily from only three, zero-order correlations. Note also that the four analysis of variance methods, in toto, do not exhaust all of the part and partial correlations one can compute. If, and only if, the independent variables and their interaction are mutually orthogonal are the values in the omega-squared column identical, but this identity does not carry over to epsilon-squared.

The advantages of correlational analysis are compounded if more than three variables are under investigation, for example, three independent variables and one dependent variable. The analyst can hold constant as needed any number of independent variables and their interactions, while many analyses of variance would be required to obtain the same information.

There are other important advantages of correlational analysis. For one, the analysis of variance not infrequently produces in both the investigator and his audience the illusion that he has experimental control over the independent variables. Nothing could be more wrong (see Underwood, 1957). Correlational analysis is less apt to produce this illusion. Furthermore, in correlational analysis, each variable in turn can be treated as if it were the dependent variable. This also deters too ready attribution of effect to cause.

Another important advantage of using correlational analysis on continuous variables is that dichotomization reduces the size of linear relationships. Statistical tests based upon continuous distributions are more powerful than those based upon dichotomized distributions. If regressions are linear, correlations among continuous variables are reduced by $y/\sqrt{pq}$, where y is the ordinate at the point of dichotomization and when the variables are dichotomized without discard of cases.

If relationships among variables are nonlinear, this information is lost entirely by the reduction of data points to two. Retention of three or four categories rather than two still provides only the roughest basis for estimating the shape of a curve. When

curvilinear regressions cannot be corrected by a transformation of the units of measurement, complications are introduced into correlational analysis; but these complications are not solved by ignoring them.

Possible interactions among variables can be handled almost as readily when working with continuous variables as with dichotomous ones. In a 2 × 2 table, the interaction is a direct function of the product of the dichotomous measures; for continuous scores the simple product will also ordinarily suffice, though more complex expressions are possible.

Computer Simulation

It was hypothesized that dichotomization might introduce spurious interactions between independent variables on the basis of the known effect on the shape of the regression of a continuous dependent variable on a dichotomized independent variable. Computer simulation was the only feasible way to attack this problem. Computer simulation also clarifies the relationship between use of equal *N*s, or unweighted means, and the other analytical analysis of variance methods.

Design of the Simulation

The 6 three-variable correlation matrices shown in Table 2 were selected so that partial correlations would be positive, zero, and negative; correlations between independent variables would be both large and small; and correlations between the two independent variables and the dependent variable would vary.

Each two-way distribution was bivariate normal in the population, and regressions were linear. Statistics based on random samples of 10,000 cases were obtained for each set of population values. Therefore, sample statistics should closely approximate population parameters. In each matrix variable, *X* was designated as the dependent variable and the other two were dichotomized at −1.0, −.5, .0, +.5, and +1.0 standard score units. Printouts were obtained of the means of the dependent variable in each of the four cells in the 25 combinations of the two dichotomies for each of the six samples. Three analysis of variance summary tables were also obtained: Methods 2 and 4, previously discussed, and the unweighted, or equal-weighted, means analysis. For the latter, 2,500 cases were assigned to each of the four cells, which is in line with the logic of the method.

Equal N *Method*

In Table 3 are shown selected data from the simulation. Correlations involving Variable *B* were selected because holding Variable *A* constant in relationships involving Variable *B* has a bigger effect than the reverse procedure. It is seen that the statistics for the equal *N* method are almost identical with those for Method 2 of Overall and Spiegel (1969) only in the data set involving a zero correlation between the two independent variables and dichotomization at the population means. If dichotomization had been at the sample medians, the results would have been identical. Retaining dichotomization at the means but allowing the correlation between the independent variables to increase (reading from left to right in the table) produces an increasing amount of overestimation of the main effect.

The reason for this increasing inaccuracy of the equal *N* procedure is that within-group variances are not homogeneous. No matter how *N* and s^2 for the four groups

TABLE 2

Population Correlations Used in the Simulation

Variable	1		2		3		4		5		6	
	A	B	A	B	A	B	A	B	A	B	A	B
X	.6	.4	.5	.4	.7	.4	.5	.2	.8	.4	.8	.2
A		.0		.5		.4		.4		.6		.7

TABLE 3
EPSILON FOR VARIABLE *B* OBTAINED BY THREE METHODS IN SIX DATA SETS FOR TWO COMBINATIONS OF DICHOTOMIES

Method	Data set					
	1	2	3	4	5	6
Dichotomization at means						
Method 2 ($r_{XB \cdot A}$)	.359	.220	.212	.053	.074	−.221
Equal *N*s	.355	.235	.220	.055	.082	−.265
Method 4 (r_{XB})	.359	.342	.356	.162	.373	.210
Dichotomization at −1.0 and −1.0						
Method 2 ($r_{XB \cdot A}$)	.287	.208	.195	.069	.089	−.127
Equal *N*s	.404	.275	.261	.062	.104	−.254
Method 4 (r_{XB})	.281	.283	.295	.127	.296	.146

vary, however, each variance is given equal weight in the pooling procedure. Furthermore, the smaller variances are typically based upon the smaller *N*s, so that there is a consistent bias downward of the estimated error variances. When dichotomization is at the means, the off-diagonal quadrants (as defined by the positive correlation between the independent variables) have smaller variances and smaller *N*s than the diagonal quadrants. Thus, the main effects in the equal *N* procedure with dichotomization at the means are the equivalent of partial correlations computed with estimates of error containing a negative bias. Thus, the partial correlations are overestimated.

For all levels of correlation between independent variables, the approximation also becomes more inaccurate as the point of dichotomization departs from the mean, although only one extreme point of dichotomization is shown in the table. Again, the direction of the error is one of overestimation of the partial correlation. The simulation involved giving an extreme 15% group the same weight as the remaining 85%; under these circumstances, both the numerator and the denominator of the correlations are distorted. Note, however, that giving groups of this sort equal weight is well within the limits of practice in the use of the pseudo-orthogonal design. A common example that truly involves an underlying continuum, though possibly not a normal one—an example which is not immediately recognized as such—is the comparison of equally weighted minority and majority groups on some dependent variable.

If both the size and sign of epsilon are recorded in all data sets for the three methods used in the simulation, it is quite clear that the equal *N* method provides estimates of main effects that approximate Method 2 much more closely than the weighting-by-*N* procedure. The equal *N* procedure does indeed lead to estimates of main effects and of interactions that approximate partial correlations rather than zero-order relationships as claimed by Humphreys and Dachler (1969a, 1969b), but the error in the approximation becomes quite large in extreme dichotomies.

The above must not be interpreted as meaning that the partial correlations approximated are necessarily those presented in Table 1 for Method 2. Theoretical relationships have not been worked out. In the present data sets, differences between Methods 1 and 2 are minimal because the population from which these samples were drawn had no interaction between the two independent variables, and interactions in the dichotomized samples were small even when nonzero.

Nonzero Interactions Produced by Dichotomization

Spurious interactions are produced by dichotomization in the absence of interaction between the continuous variables. These are quite small in Method 2, but their consistency supports the generalization. The zero-order interactions in Method 4 can be quite large, but these of course reflect in large part the size of main effects. To assess Method 4 interactions, it is necessary to compute partial correlations.

Interactions obtained by the equal-weighting procedure are perhaps the only ones that might be considered psychologically significant and which might be statistically significant in much smaller samples. This is seen in Table 4, which contains the mean and maximum absolute values of epsilon for each data set. The largest means are associated with data sets involving the higher intercorrelations, but

TABLE 4
MEANS AND MAXIMA OF EPSILON FOR INTERACTION IN SIX DATA SETS FOR ALL DICHOTOMIES

Item	Data set					
	1	2	3	4	5	6
Method 2						
M	.000	.018	.027	.030	.046	.060
Maximum	.000	.041	.055	.049	.076	.092
Equal *N*						
M	.000	.037	.045	.050	.101	.183
Maximum	.011	.106	.084	.090	.180	.336

the heterogeneity in size of these correlations is also involved. Maximum values appear in combinations involving extreme dichotomies.

Numerically Small, But Seemingly Large, Interactions

Crossovers of the sort reported by Jensen (1968), in which a difference in means at one level of an independent variable is different in sign from the difference at another level, can occur even though the size of the interaction using appropriate methodology is quite small. To those accustomed to the use of the analysis of variance with orthogonal data this seems quite incongruous since a crossover is, under those circumstances, evidence for an interaction. For correlated independent variables, however, the crossover represents rather closely the linear expectation, subject only to the small spurious interaction previously described.

A necessary, but not a sufficient, condition for a single crossover to occur is a small though still positive, partial point biserial for one of the independent variables. It is also necessary that the more highly correlated independent variable be represented by an extreme dichotomy. The crossing over is, of course, independent of whether an unweighted analysis or a least squares analysis is used, but the crossing over is larger and more likely to be statistically significant in the unweighted analysis.

Double crossovers, in which the difference in means at both levels of one of the independent variables has the same sign but differs in sign with respect to the weighted by *N* main effect, also occur. To casual observation this suggests some sort of complex interaction, but it again represents the linear expectation with relatively little error. For double crossovers to occur, the partial correlation for one of the independent variables with the dependent variable must be negative.

It is not possible to summarize all of the relevant computer output, but examples have been selected of both single and double crossovers and are presented in Table 5.

DISCUSSION AND CONCLUSIONS

There are warnings by highly respected authorities concerning arbitrary orthogonalization of correlated independent variables. Thus, McNemar (1969) has a section (pp. 444–449) concerning this problem, and Winer (1962) draws the following conclusion:

> On the other hand, should the cell frequencies be directly related to the size of corresponding population strata, then such frequencies should be used in estimating the mean of the population composed of such strata [p. 224].

This advice has not only been disregarded theoretically, as in the paper by Fischbach and Walberg (1971), but it has also been disregarded with increasing frequency in data analysis by research workers. Educational psychology, child development, and

TABLE 5
*N*S AND MEANS FOR SUBGROUPS AND FOR MAIN EFFECTS IN TWO DATA SETS ILLUSTRATING SINGLE AND DOUBLE CROSSOVERS

Number/ mean	Data set 5			Data set 5		
n	849	7517	8366	3583	4783	8366
M	−.21	.52	.45	.03	.76	.45
n	752	882	1634	1379	255	1634
M	−2.41	−2.49	−2.45	−2.44	−2.51	−2.45
n	1601	8399		4962	5038	
M	−1.24	.21		−.65	.59	
	Data set 6			Data set 6		
n	775	7650	8425	1247	3704	4951
M	.85	.42	.46	1.89	1.09	1.29
n	811	764	1575	3753	1296	5049
M	−1.94	−3.02	−2.46	−1.06	−1.85	−1.27
n	1586	8414		5000	5000	
M	−.58	.11		−.33	.33	

personality research perhaps head the list in frequency usage of the pseudo-orthogonal design, but traditional experimental fields are not pure in this regard. An aura of the laboratory and of controlled experimentation accompanies the use of the analysis of variance, while correlational analysis is perceived as not experimental. Every undergraduate, or almost every one, knows that correlation does not mean causation. It is not as well known that t ratios and F ratios are also measures of relationship and that it is the ability of the experimenter to make random assignments of persons to treatment groups that distinguishes experimental control from statistical control. To put it bluntly, the analysis of variance has become more popular by far than it deserves, and not only in research on individual differences. The preferred alternative varies from one set of problems to another, but for most research involving individual-differences variables, correlational analysis is strongly preferred to the analysis of variance.

Some of the literature involving the use of the pseudo-orthogonal design can be salvaged. Interpretation of main effects as approximations to partial correlations rather than zero-order relationships is legitimate and useful. If dichotomization is at the means or medians, the approximation is reasonably accurate. If the dichotomies in the population are of grossly unequal size, however, the amount of error in the approximation is substantial.

With respect to the interactions stemming from the use of the pseudo-orthogonal design, one cannot be as sanguine about salvage. These must be analyzed carefully with respect to possible spuriousness. With correlated variables, a "crossover" does not in itself support the interpretation of interaction. Use of products of continuous measures in correlational analysis is legitimate and should be encouraged. This is especially true for selected products dictated by theory or even hunch. On a trial-and-error basis, on the other hand, product terms could easily exhaust available degrees of freedom.

There are still other cases in which the literature as a whole should probably be discarded and a fresh start should be made. Consider such designs as the one employed by Orn and Das (1972). Socioeconomic status and intelligence were made orthogonal by equating Ns. (The authors were engaged in a replication of some of Jensen's, 1968, research.) In addition, however, groups were equated, more or less, for mental age. This, in turn, insured large differences between IQ groups in chronological age. It is significant that only about 9% of the children surveyed could be selected for the experimental groups. With one dependent rote-learning variable, they replicated Jensen's finding of a crossover involving socioeconomic status and IQ, though failing to note that chronological age and socioeconomic status also, of necessity, interacted. In such data, only the fact of replicability is of any importance since psychological interpretation is impossible, or virtually so. Prime candidates for the research wastebasket are pseudo-orthogonal designs that also involve matching on still other individual-differences measures.

REFERENCES

Fischbach, T. J., & Walberg, H. J. Weighted and unweighted means for estimation: A note on the Humphreys-Dachler and Jensen papers. *Journal of Educational Psychology*, 1971, **62**, 79–80.

Hays, W. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.

Humphreys, L. G., & Dachler, H. P. Jensen's theory of intelligence. *Journal of Educational Psychology*, 1969, **60**, 419–426. (a)

Humphreys, L. G., & Dachler, H. P. Jensen's theory of intelligence: A rebuttal. *Journal of Educational Psychology*, 1969, **60**, 432–433. (b)

Jensen, A. R. Patterns of mental ability and socioeconomic status. *Proceedings of the National Academy of Sciences*, 1968, **60**, 1330–1337.

McNemar, Q. *Psychological statistics*. (4th ed.) New York: Wiley, 1969.

Orn, D., & Das, J. Socioeconomic status and short-term memory. *Journal of Educational Psychology*, 1972, **63**, 327–333.

Overall, J. E., & Spiegel, D. K. Concerning least squares analysis of experimental data. *Psychological Bulletin*, 1969, **72**, 311–322.

Peters, C., & Van Voorhis, W. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.

Underwood, B. *Psychological research*. New York: Appleton-Century Crofts, 1957.

Winer, B. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received March 2, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 473-480

REINFORCEMENT PRACTICES OF BLACK AND WHITE TEACHERS IN INTEGRATED CLASSROOMS[1]

ROBERT BYALICK
Center for Community Adjustment, Farmingdale, New York

DONALD N. BERSOFF[2]
University of Georgia

The Positive Reinforcement Observation Schedule was used to obtain data concerning expressed preference for and observed use of positive reinforcement in integrated classrooms taught by 30 black and 30 white female elementary school teachers. Results indicated the following: (*a*) Black and white teachers emit virtually equal rates and types of reinforcers. (*b*) Only 32% of the combined sample actually use the reinforcers they state they prefer. (*c*) Reinforcement emission in classrooms occurs at a relatively low rate. (*d*) Teachers reinforce opposite-raced children more frequently than children of their own race. (*e*) Males are reinforced more frequently than females. (*f*) Black females are the least reinforced group of all. These results are related to previous research and implications are drawn.

It is now well established that contingent reinforcement by teachers has direct effects on both academic and nonacademic behavior of students. The interaction of variables such as age (Cradler & Goodwin, 1971; Rosenhan & Greenwald, 1965), sex (Rosenhan & Greenwald, 1965; Rucinski, 1968) intelligence (Wolfensberger, 1960), and social class (Cradler & Goodwin, 1971; Shores, 1969; Zigler & Kanzer, 1962) as they influence reinforcement effectiveness has also been widely studied (though with few findings that are either clear or consistent). Yet, Staats' (1968) indictment that systems of reinforcement in school situations have never been subjected to systematic study and research remains correct. With very few exceptions (e.g., Meyer & Lindstrom, 1969), observations of reinforcing behavior in the natural setting of the classroom are absent from the literature. Thus, while there is a genuine concern over the propriety and utility of concrete, extrinsic and intrinsic reinforcement as they relate to children's learning, there is a dearth of normative data regarding the actual emission of the variety of reinforcers available to teachers.

If normative data concerning teacher reinforcement emission is scanty, such information is almost nonexistent for biracial samples of teachers. Increasing numbers of black and white teachers are facing integrated classrooms, yet very little empirical data concerning teacher–child interaction in such classes has been published. Reviewing the literature recently, Meyer and Lindstrom (1969) commented that observational studies of teacher behavior toward black and white children were unknown to them. Their own study, comparing the rates of approval and disapproval in biracial classrooms, yielded no differences between black and white teachers toward classes of mixed race/sex children. However, only 13 teachers were studied and only 4 of those were black.

Lastly, previous research (Bersoff & Moyer, 1973) has suggested that teachers have rather clear ideas concerning the kinds of reinforcers they believe to be the most effective in the classroom. What is not known is whether teachers actually use those reinforcers they state they prefer. In

[1] Some of the data presented originally appeared in the senior author's doctoral dissertation completed at the University of Georgia.

[2] Reprints may be requested from Donald N. Bersoff, who is now at the University of Virginia Law School, Charlottesville, Virginia 22902.

the light of the above concerns, the present investigation had three major goals:

1. Begin a systematic inquiry into the reinforcement practices of teachers as they exhibit them under natural classroom conditions.

2. Compare the rate of reinforcement emission of a sizable sample of black and white teachers in integrated classrooms.

3. Compare the rates of actual reinforcement emission of teachers with their stated preferences.

Method

Subjects

The subjects were 60 female elementary school teachers, 30 white and 30 black, from 15 schools in the southeastern United States. Teachers were selected based on their willingness to involve themselves in the study. The nature of the study was explained as the researchers' desire to observe teacher–child interaction. Five white teachers and five black teachers from each of Grades 1 through 6 participated. Each group had equivalent teaching experience (white teachers = 14.6 years; black teachers = 15.9 years) and an equal number in each group (9 out of 30) had had direct exposure to behavioral technology.

The classes of the participating teachers included a total of 1,398 children. Table 1 categorizes the number of children by sex and race for the two groups of teachers. Only those classrooms with at least 20% of a single race were observed.

Instruments

The Positive Reinforcement Observation Schedule was the instrument used to observe and record reinforcement emission. Extended descriptions of the scale, including validity and reliability data, have been reported elsewhere (Bersoff, 1973; Bersoff & Moyer, 1973). Briefly, the scale consists of 10 categories of behaviors that may be emitted by mediators (i.e., teachers, parents). These behaviors appear to possess reinforcing qualities that are both powerful and durable, but there is no assumption made concerning the trans-situational nature or generalized effectiveness of these categories. Underlying the Positive Reinforcement Observation Schedule is the concept that reinforcement is a relational rather than an absolute property of any activity. The scale may be used as a mediator preference scale when constructed in a paired-comparison format and as an observation schedule to obtain frequency and rate of positive reinforcement. The 10 positive reinforcement categories, their symbols, and highly abbreviated definitions follow:[3]

Administration of Concrete Rewards, Direct (CRD): giving immediate concrete rewards such as candy, money, or free time;

Administration of Concrete Rewards, Token (CRT): giving symbolic rewards redeemable for direct concrete rewards at some future time;

Affirmation of Appropriate Behavior (AAB): verbal contact indicating that responses are correct, acceptable, or appropriate;

Rapport–Praise (RP): evaluative reactions which go beyond the level of simple affirmation. Rapport–Praise communicates a positive evaluation or a warm personal reaction;

Positive Facial Attention (FA+): looking at target when mediator is smiling or attending to what target is doing or saying;

Positive Physical Contact (PC+): patting, embracing, holding arm, taking hand;

Accepts Feelings (AF): accepting or clarifying feeling tone of target in nonthreatening manner;

Accepts Ideas (AI): clarifying, building, or developing ideas suggested by target;

Adjuvant Mastery (AM): urging, prompting, fostering, promoting confidence, success, providing encouragement for response production;

Aiding by Example (AE): demonstration of appropriate behavior by mediator when target is either nonresponsive or incorrect in exhibiting expected response.

Procedure

The subjects were observed while teaching a lesson involving the entire class. Observers sat at the side or back of the classroom to minimize interference with the natural classroom environment and to better view any nonverbal teacher behavior. No observations were made on Monday mornings, Friday afternoons, or immediately before or after holidays. Observers entered the classroom approximately one half hour before the 45-minute lesson began. Actual recording did not begin until the last half hour of the lesson. Thus, data collection did not begin for at least 45 minutes after the observer entered the room. The positive reinforcement behaviors which were observed and tallied were those corresponding to Positive Reinforcement Observation Schedule events.

Observations were made by the senior author and by another trained observer, a female graduate student in educational psychology with prior experience as a teacher.[4] The senior author made 60% of the observations (19 out of 30 in classes with black teachers and 17 out of 30 in classes with white teachers).

Prior to data collection, three teachers—one black and two white—were observed to obtain acceptable interjudge agreement. The two observers were seated on opposite sides of the room and upon signal, teacher behavior was recorded for 30 minutes. Only those teacher behaviors which followed

[3] Complete definitions with behavioral examples may be secured from the second author.

[4] The authors wish to express their appreciation for the work of Maggie Weshner who served as the second observer.

TABLE 1

OBSERVED USE AND EXPRESSED PREFERENCE OF POSITIVE REINFORCEMENT IN BIRACIAL CLASSROOMS

PROS event	No. teachers using event	Frequency of use with children					Observed rank	% used	Expressed rank
		White		Black		Total			
		Males	Females	Males	Females				
Black teachers (n = 30)[a]									
n		244	200	115	120	679			
AAB	30	306	133	93	54	585	1	.471	2
FA+	27	117	61	38	27	243	2	.195	3
AI	26	86	29	26	14	115	3	.125	4
AF	22	33	15	5	7	60	4	.048	6
AE	16	28	15	9	4	56	5	.045	9
PC+	15	23	13	10	8	54	6	.043	7
RP	10	24	11	8	3	46	7	.037	1
AM	11	16	3	11	3	33	8	.027	5
CRT	3	4	3	1	3	11	9	.009	10
CRD	0	0	0	0	0	0	10	.000	9
Total		637	283	200	123	1243		1.000	
$\bar{X}$ positive reinforcements per child		2.61	1.42	1.74	1.02	1.83			
White teachers (n = 30)[b]									
n		254	227	131	107	719			
AAB	29	184	125	49	88	546	1	.450	2
FA+	29	93	69	80	30	272	2	.224	5
AI	22	52	53	32	15	152	3	.125	3
AF	7	2	7	1	0	10	9	.008	6
AE	16	19	14	16	8	57	5	.047	7.5
PC+	13	25	13	6	6	50	6	.041	7.5
RP	14	22	35	13	14	84	4	.069	1
AM	8	13	3	8	2	26	7	.021	4
CRT	0	0	0	0	0	0	10	.000	10
CRD	1	5	6	3	3	17	8	.014	9
Total		415	325	308	166	1214		1.000	
$\bar{X}$ positive reinforcements per child		1.63	1.43	2.35	1.55	1.69			

Note. Abbreviations: PROS = Positive Reinforcement Observation Schedule; AAB = Affirmation of Appropriate Behavior; FA+ = Positive Facial Attention; AI = Accepts Ideas; AF = Accepts Feelings; AE = Aiding by Example; PC+ = Positive Physical Contact; RP = Rapport–Praise; AM = Adjuvant Mastery; CRT = Administration of Concrete Rewards, Token; and CRD = Administration of Concrete Rewards, Direct.

[a] $\bar{X}$ positive reinforcements per teacher = 41.43 and $\bar{X}$ positive reinforcements per teacher per minute = 1.38.

[b] $\bar{X}$ positive reinforcements per teacher = 40.46 and $\bar{X}$ positive reinforcements per teacher per minute = 1.35.

appropriate student response were considered positive reinforcements. The three observation sessions yielded over 150 Positive Reinforcement Observation Schedule events. A Pearson product-moment correlation between observers for the frequencies of the 10 Positive Reinforcement Observation Schedule categories was calculated for each teacher observed. Coefficients were uniformly high (.97 for Teachers 1 and 2; .99 for Teacher 3). Observations for the purpose of hypotheses testing were conducted by single observers. However, as a continuing check on observer agreement, two additional joint observations were made during the data collection period. Pearson product-moment correlations for these two observations were .96 and .95 (average r for all five agreement observations = .974).

In calculating reliability, some Positive Rein-

forcement Observation Schedule categories were not used by teachers, possibly inflating the correlation. As a more rigorous check, correlations were calculated excluding the categories with zero frequencies. Results indicated that the agreement between observers remained high. Recalculated correlations ranged from .882 to .999 (average r = .963).

To obtain positive reinforcement preferences, after the observation period was completed, each teacher was administered the Positive Reinforcement Observation Schedule in its paired-comparison form. The teachers were instructed to determine which of the listed positive reinforcement categories they felt would be the most potent reinforcers with children. In filling out the scale, teachers were given the following definition: Positive reinforcement is behavior by a teacher following student response for the purpose of strengthening or accelerating appropriate behavior.[5]

Results

Table 1 presents a summary of the data used to answer the following questions:

1. Do black teachers and white teachers emit differential rates of reinforcement?
2. Do black teachers and white teachers differ in the kinds of reinforcing events they emit?
3. Do black teachers and white teachers differentially reinforce students of a certain sex, race, or combination of both?
4. Do teachers emit the kinds of reinforcement they express a preference for?

Amount of Reinforcement

Frequency of positive reinforcement emission was virtually equal for both races (black teachers, M rate = 1.38 positive reinforcements per minute, SD = .456; white teachers, M rate = 1.35 positive reinforcements per minute, SD = .493; t = .303, df = 58, $p > .05$). When rate/child was examined, differences remained negligible (black teachers = 1.83; white teachers = 1.69).

Kinds of Reinforcement

The types of positive reinforcements black teachers and white teachers used in the classroom were very similar. When a rank-order correlation for observed use between the two groups was calculated, the result was a rho of .758 ($p < .01$). When frequencies rather than ranks were considered, the resultant Pearson-product-moment correlation was even higher (r = .90, $p < .001$). For both groups, AAB, FA+, and AI were the most frequently used categories, comprising almost 80% of all reinforcement emitted in classrooms. The groups differed by three or more ranks for only two categories. Black teachers accepted student feelings (AF) with greater frequency while white teachers emitted more warm evaluative verbal comment (RP).

Reinforcement by Race of Teacher and Race/Sex of Child

This question was investigated by calculating a t test for the contrast corresponding to the triple-order interaction of race of teacher and race/sex of child using a procedure adapted from Wiley (1970). To adjust for the different numbers of male/female and black/white children in each class, corrected scores were used. These scores were obtained by dividing the number of children in each of the four groups (white males, white females, black males, black females) into the number of positive reinforcements given to each group by a particular teacher (e.g., 10 black males receive 20 positive reinforcements, corrected score = 2). To obtain the values necessary for calculating t, the sum of corrected scores of white females and black males was subtracted from the sum of corrected scores of white males and black females. This method permitted the examination of the contrast between the differences of the various groups.

Using such a procedure, the overall t calculated was 1.67 (df = 58) falling somewhat short of the t (t = 2.00) needed for significance. However, the result was close enough to encourage examination of the interaction within each teacher group. Thus, t tests contrasting the differences among corrected scores of the various groups were calculated.

Among the six contrasted groups, black teachers significantly reinforced one group in each of four pairs more often than the

[5] Directions and a copy of the original paired-comparison format are available from the second author.

other, while white teachers did the same for one group in each of two contrasted pairs. Specifically, the following findings resulted from the triple-order interaction process.

From black teachers:

White males received more positive reinforcement than black females ($t = 5.50$, $df = 29$, $p < .001$).

White males received more positive reinforcement than white females ($t = 4.90$, $df = 29$, $p < .001$).

White males received more positive reinforcement than black males ($t = 3.00$, $df = 29$, $p < .01$).

Black males received more positive reinforcement than black females ($t = 2.50$, $df = 29$, $p < .02$).

Differences between white females and black males were insignificant ($t = 1.07$, $df = 29$).

Differences between white females and black females were insignificant ($t = 1.07$, $df = 29$).

From white teachers:

Black males received more positive reinforcement than white females ($t = 2.10$, $df = 29$, $p < .05$).

Black males received more positive reinforcement than white males ($t = 2.05$, $df = 29$, $p < .05$).

Differences between black males and black females were insignificant ($t = 1.37$, $df = 29$).

Differences between white males and white females were insignificant ($t = 1.20$, $df = 29$).

Differences between white females and black females were insignificant ($t = 0.87$, $df = 29$).

Differences between white males and black females were insignificant ($t = 0.10$, $df = 29$).

Therefore, teachers (female) of both races bestowed significantly greater amounts of positive reinforcement on male than on female children, though this was a more uniform finding among black teachers. Second, children of the race opposite to that of the teacher received more positive reinforcement than did same-raced children. The amount was significantly greater in four of the eight contrasts possible in this regard.

Expressed Preference and Observed Use of Positive Reinforcement

Rank-order correlations were calculated between each teacher's observed use of positive reinforcement in the classroom and the preferences she expressed on the paired-comparison format of the Positive Reinforcement Observation Schedule. For 40% of the white teachers (12 out of 30), there was a significant positive relationship between expressed preference and actual use of Positive Reinforcement Observation Schedule events. In contrast, 23% (7 out of 30) of the black teachers showed such a significant relationship. Thus, less than 32% of all teachers' ranked use of positive reinforcement matched ranked expression of effectiveness. Table 2 provides the frequency distribution of rhos for each group of teachers. Black teachers had fewer scores in the range of significance and fewer scores approaching significance.

By combining data, rankings were obtained for expressed preference and observed use for each racial group. Rho correlations between such rankings were calculated. Results indicated that significant cor-

TABLE 2

FREQUENCY DISTRIBUTION OF RANK-ORDER CORRELATIONS BETWEEN OBSERVED USE AND EXPRESSED PREFERENCE OF POSITIVE REINFORCEMENT FOR WHITE TEACHERS AND BLACK TEACHERS

ρ[a]	Teachers	
	White	Black
.90–1.00	0	0
.80–.89	3	0
.70–.79	2	2
.60–.69	5	5
.50–.59	4	1
.40–.49	3	4
.30–.39	2	2
.20–.29	5	5
.10–.19	0	5
.00–.09	1	2
−.10–.00	5	4

[a] $r' = .564$ ($p < .05$).

relations exist within both groups (white teachers, $\rho = .718$; black teachers, $\rho = 6.12$; $p < .05$). Reference to Table 1 shows where the commonalities and disparities between preference and use lie. While the group correlations are high, the individual correlations seem to present a more accurate and realistic view. Most teachers do not use the kinds of reinforcers they select as the most potent in accelerating children's appropriate behavior.

One further comparison provides further data as to the overall commonality between black teachers and white teachers. When expressed preferences are ranked and compared, the relationship between the groups becomes almost uniformly congruent ($\rho = .961$).

Discussion

The present investigation clearly indicates that black and white teachers are virtually mirror images of each other when amount and kind of reinforcement emission is the matter of concern. These results may lay to rest some fears concerning the performance of black teachers and white teachers in integrated classrooms. Yet, when teacher reinforcement practices are viewed as a whole, there are some disheartening findings. Despite repeated evidence that many children, particularly those of low socioeconomic status, perform better when given material rewards (e.g., Benowitz & Busse, 1970; Cradler & Goodwin, 1971; Terrell, Durkin, & Wiesley, 1959; Zigler & Kanzer, 1962), both black and white teachers in this study gave very few concrete reinforcements. Only 28 CRD or CRT were emitted of over 2,400 observed Positive Reinforcement Observation Schedule events. Only 4 teachers out of 60 used these categories, and 1 of those 4 teachers accounted for 17 of the 28 events noted. Miniscule use of CRD and CRT matched expressed preference. Both groups of teachers thought that concrete rewards were the two least potent means for accelerating appropriate behavior in children.

Further, teachers strongly preferred some of the positive reinforcement categories, although they hardly used them. For example, both groups ranked AM in the top five during the expressed preference task but used it less than 3% of the time (observed use rank, black teachers = 8, white teachers = seven). Rapport–Praise was the overwhelming choice as the most effective positive reinforcement, but it was used less than 6% of the time. Thus, while the data indicated that few differences exist between black teachers and white teachers in positive reinforcement emission, it may also be said that teachers as a whole do not act in accordance with their individual preferences. In the majority of cases, teachers did not use those reinforcers they thought to be the most potent.

A subjective interpretation of the data concerning observation of positive reinforcement emission in classrooms is that teachers are more interested in reinforcing academic activity in traditional ways, mainly through what might be called the "distant reinforcers" (AAB, FA+, AI) than through "proximity reinforcers" (CRD, CRT, PC+) involving material rewards and close personal contact. This hypothesis is corroborated by both previous research investigating expressed preferences only (Bersoff, 1974)[6] and the present study, during which it was noted that the majority of teachers maintained physical distance from their students and reinforced without moving from behind their desks.

A final disappointing note concerning reinforcement emission is that on the average, each child received about 1.75 positive reinforcements per 30-minute observation period. With time out for a wide variety of nonacademic activity, one could be generous and assign four hours of the school day to academic work. Thus, each child is the recipient of about 14 positive reinforcements per academic day. Not enough is known about schedules of reinforcement in the acceleration and maintenance of human classroom learning, but this amount would seem to be a rather lean schedule for the

[6] Further data is provided in a manuscript entitled "Disparities in reinforcement preferences between mediators and modifiers" by the second author which has been submitted for publication.

production of academic behavior. It is reasonable to assume that teachers are capable of reinforcing at higher rates. When one considers Samph's (1968) finding that teachers tend to inhibit criticism when aware of observers' presence, the data gathered in the present study may represent the upper limits (without other kinds of intervention) of positive teacher behavior.

Our main finding, that no substantial differences in amount and kind of positive reinforcement occurred between black teachers and white teachers in the sample investigated, supports the work of Meyer and Lindstrom (1969), who studied similar teacher behavior in Head Start classrooms. They concluded that "there [were] no real differences among teachers in the frequency with which they give approval ... to their youngsters [p. 24]." However, they also found that their sample of black teachers and white teachers distributed approval relatively evenly among the race and sex groupings of the children observed. That finding was certainly not corroborated by our study.

The investigation of the effects of race and sex and positive reinforcement emission are most enlightening, revealing something akin to reverse prejudice. Black teachers reinforce white children more frequently than white teachers, and white teachers do just the opposite. Most of the differences are accounted for by the fact that teachers of both races reinforced opposite-raced boys significantly more than same-raced or same-sexed children; that is, teachers reinforced males of the opposite race most frequently in comparison to other groupings. The subgroup receiving the next highest amount of positive reinforcement was males of the same race as the teacher. Both black teachers and white teachers gave more positive reinforcement to females of the opposite race than to the same race as themselves. The group least reinforced were black females taught by black teachers.

The racial differences are in line with previous research (Brown, Payne, Lankewich, & Cornell, 1970) which found that teachers gave more praise and less criticism to children of the opposite race. Brown et al. (1970) studied only fully segregated classrooms. However, results of the present study may lend themselves to developing a generalization that children of the race opposite to that of the teacher are interacted with more favorably than children of the same race.

The sex differences are not so readily comprehended. Baughman (1971) noted that black females outperform black males and closely match white males on many academic and intellectual measures. Thus, it is somewhat surprising that black females receive much fewer teacher reinforcements than any other subgroup. This finding is most clearly evident in classes taught by black teachers. It is interesting to note in this regard that similar conditions are found in adult life where the black, female-headed family is at the bottom of the economic structure.

In a final elaboration of the data, which we offer with great tentativeness, there is one small sign of racial biasing by black and white teachers. Although teachers tend to reinforce opposite-raced children more frequently, this is not true within the category of PC+. White teachers touch white children with greater relative frequency than black children, and black teachers touch black children with greater frequency. The difference is most pronounced with regard to black males and white males in groups taught by white teachers. Out of the pool of reinforcement received, PC+ was administered 6% of the time to white males and only 2% of the time to black males in classes taught by white teachers. It was only within the PC+ category that such a reversal of reinforcement frequency was found.

It should be noted, in conclusion, that the sample of teachers observed was all female. None of these results may be germane to male elementary teachers. But it would be especially interesting to see if females were generally neglected by all teachers, or if this phenomenon was sex specific. In any case, as most teachers in the elementary grades are female, and assuming our findings are replicable, whatever prejudice

women are experiencing may be a consequence of "doing unto oneself."

REFERENCES

BAUGHMAN, E. E. *Black Americans: A psychological analysis.* New York: Academic Press, 1971.

BENOWITZ, M. L., & BUSSE, T. V. Material incentives and the learning of spelling words in a typical school situation. *Journal of Educational Psychology,* 1970, **61,** 24–26.

BERSOFF, D. N. Behavioral approaches to assessment and observation in the school. In J. F. Magary (Ed.), *Handbook of school psychological services.* St. Louis, Missouri: C. V. Mosby, 1974.

BERSOFF, D. N., & MOYER, D. Positive Reinforcement Observation Schedule (PROS): Development and applications to educational settings. *Proceedings of the 81st Annual Convention of the American Psychological Association,* 1973, **8,** 713–714. (Summary)

BROWN, W. E., PAYNE, L. T., LANKEWICH, L., & CORNELL, L. L. Praise, criticism, and race. *Elementary School Journal,* 1970, **70,** 373–377.

CRADLER, J. D., & GOODWIN, D. L. Conditioning of verbal behavior as a function of age, social class, and type of reinforcement. *Journal of Educational Psychology,* 1971, **62,** 279–284.

MEYER, W. J., & LINDSTROM, D. The distribution of teacher approval and disapproval of Head Start children. Syracuse, N. Y.: Syracuse University, Evaluation and Research Center, 1969. (Mimeo)

ROSENHAN, D. L., & GREENWALD, J. A. The effects of age, sex, and socioeconomic class on responsiveness to two classes of verbal reinforcement. *Journal of Personality,* 1965, **33,** 108–121.

RUCINSKI, P. R. The motivating effect of two reinforcers upon lower- and middle-class fifth-grade children. *Journal of Educational Research,* 1968, **61,** 368–371.

SAMPH, T. Observer effects on teacher behavior. Unpublished doctoral dissertation, University of Michigan, 1968.

SHORES, R. Performance of learning disabled and normal children from differing social classes under tangible or intangible reinforcement conditions. *Exceptional Children,* 1969, **35,** 643–644.

STAATS, A. W. *Learning, language, and cognition.* New York: Holt, Rinehart & Winston, 1968.

TERRELL, G., JR., DURKIN, K., & WIESLEY, M. Social class and the nature of the incentive in discrimination learning. *Journal of Abnormal and Social Psychology,* 1959, **59,** 270–272.

WILEY, D. E. Design and analysis of evaluation studies. In M. C. Wittrock & D. E. Wiley (Eds.), *The evaluation of instruction: Issues and problems.* New York: Holt, Rinehart & Winston, 1970.

WOLFENSBERGER, W. Differential rewards as motivating factors in mental deficiency research. *American Journal of Mental Deficiency,* 1960, **64,** 902–906.

ZIGLER, E., & KANZER, P. The effectiveness of two classes of verbal reinforcers on the performance of middle and lower class children. *Journal of Personality,* 1962, **30,** 157–163.

(Received April 6, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 481-485

INCIDENTAL AND RELEVANT LEARNING WITH INSTRUCTIONAL OBJECTIVES[1]

PHILIPPE C. DUCHASTEL[2] AND BOBBY R. BROWN[3]

Florida State University

It was hypothesized that one role of objectives in learning is to serve as orienting stimuli by which the learner can decide which material to concentrate on and which to pay less attention to. With a brief text to learn, 58 college students received either one half of the 24 objectives for the text or no objectives at all. As expected, the subjects with half of the objectives performed better than their counterparts without objectives on the posttest items referenced to their objectives (relevant learning) and less well on the items not covered by their objectives (incidental learning). These findings conflicted with previous research results with respect to incidental learning; this could have resulted from the fact that the subjects in the present study had practical experience with an objective-referenced instructional model.

Much research has been done and is presently being conducted on the effects of instructional objectives in learning. One aspect of that research has addressed the question of whether providing students with advance knowledge of the instructional objectives for a unit of instruction facilitates their learning of that unit. A review of the results obtained in this area which has recently been completed (Duchastel & Merrill, 1973) pointed to the great variability involved in the conclusions drawn from these research efforts.

Although a number of studies have failed to support the hypothesis that students provided with objectives achieve more than students unaware of the objectives, a sufficient number of investigations have confirmed the hypothesis to lead to an affirmative opinion on the question. It would therefore seem appropriate to now view the issue on a more basic level and to investigate various reasons why objectives could possibly be helpful to students. The present study addresses one aspect of this issue, namely, that objectives facilitate student learning by providing direction for that learning.

This directive function of objectives can be viewed within the general framework of the theory evolving around the use of orienting stimuli (Rothkopf, 1970). Basically, orienting stimuli are thought to elicit inspection behaviors which in turn determine what is learned. Orienting stimuli should, therefore, focus the student's attention on the important aspects of the content (whatever is so defined as important) and minimize his attention on the incidental or illustrative parts of the learning material. This focusing effect should increase performance on test items referenced to the important aspects of the material and decrease performance on those items which are referenced to the incidental aspects.

The main findings from orienting research were that (*a*) inserting questions in reading material enhanced performance on question-relevant items in the posttest and (*b*) performance on nonrelevant items (those not

[1] Much of the research reported here was conducted while the authors were at the Computer-Assisted Instruction Center, Florida State University. The authors gratefully acknowledge the cooperation of D. Hansen, B. Kibler, F. J. King, and P. Merrill, who reviewed a draft copy of this paper.

[2] Requests for reprints should be sent to Philippe C. Duchastel, who is now at the Centre d'Application des Média Technologiques à l'Enseignement et à la Recherche (CAMTER), Université du Québec à Montréal, B.P. 3050, Succursale B, Montréal 110, Québec, Canada.

[3] Now director of the Computer-Assisted Instruction Center, University of Iowa, Iowa City.

referenced to inserted questions) was generally improved through the use of questions placed after the learning passage but was not improved through the use of prequestions. In some cases (Frase, 1968; Frase, Patrick, & Schumer, 1970; Patrick, 1968), prequestions actually depressed incidental learning. Presumably, then, questions which are placed before the material focus the student's attention on question-relevant material and not on the incidental material.

With respect to instructional objectives, Morse and Tillman (1972) gave half of their 52 subjects three of the six objectives developed for a unit of instruction. The other half received no objectives. Overall, the subjects receiving the partial list of objectives performed significantly better on the test items referenced to these objectives than on items not related to these objectives. The subjects receiving no objectives performed equally well on either set of items. Incidental learning for the group with objectives was not adversely affected.

Rothkopf and Kaplan (1972) also contrasted the effects of objectives on intentional and incidental learning. The experimental groups provided with objectives performed better on intentional than on incidental learning. However, they also performed better on incidental learning than a control group not provided with objectives who were simply told to learn "everything" in the unit.

The preceding two studies have found that objectives, while enhancing relevant learning, do not, however, depress incidental learning. This finding is somewhat unexpected and in conflict with results obtained with the use of prequestions. One possible explanation for these findings is that the subjects used in the two studies may not have been familiar enough with the role of objectives to fully use them in focusing their learning. It has been pointed out by a number of researchers (cf. Tiemann, 1968) that the possible effects of objectives may not be detected in research in which the subjects have not fully accepted the idea that the posttest which they will be taking is directly referenced to the objectives presented to them. This consideration would seem to be especially crucial in the issue with which we are presently dealing. If a student thinks that his instructor might test him on all the material and not just the material delimited by the objectives, he is likely not to focus his attention on the objectives as much as he would otherwise.

It would seem, therefore, that the ideal subjects to use in research on objectives are students who have had practical experience with objectives and criterion-referenced testing in their academic courses. The purpose of the present research effort was to investigate the incidental–relevant learning hypothesis with such a group. It was expected that not only would objectives enhance relevant learning but that they would also depress incidental learning.

Method

Subjects

A total of 58 college students participated in the study. These students were volunteers from a communication course at Florida State University and received course credit for their participation.[4] The course in question was a mastery course organized around a set of established objectives provided to the students in which each unit test was directly referenced to the unit objectives. The present study was conducted after the students had taken four unit tests so as to insure that the students were fully aware, during the experiment, of the role played by objectives in learning. The students had also had a lecture at the beginning of the course explaining to them how to proceed through the course using the objectives.

Materials

The instructional materials consisted of a slightly modified reading passage taken from a text entitled *The Mushroom Handbook* by Krieger (1967). The passage, which was 10 pages long (approximately 2,400 words), was taken from the section entitled "Conditions Under Which Mushrooms Grow and Thrive" and dealt with such aspects of development as food and temperature requirements, parasitism, fairy rings, etc. These materials were selected mainly because of their presumed unfamiliarity to the typical undergraduate student and because they seemed quite typical of much of the course material found at the college level.

The instructional objectives used in this study numbered 24 and were developed from an exami-

[4] The authors are grateful to Bob Kibler, M. Ron Basset, and M. Tom Porter, who made the course available for research purposes and offered valuable suggestions for the study.

nation of the passage. All objectives stated what the student would be expected to do once he finished studying the text. All of the objectives were very specific and were related to the knowledge category of Bloom's (1956) taxonomy of objectives. The following 2 objectives were typical of those presented: [After completing this unit, you will be expected to] (*a*) give two examples of plants which form a cooperative symbiosis with fungi and (*b*) state the name for a plant's response to gravity.

The posttest was developed so as to reflect directly the instructional objectives. One item was written for each objective, for a total of 24 items. All items were of a constructed response format which tapped recall rather than recognition. The items referenced to the two objectives presented above were as follows: (*a*) Give two examples of plants which form a cooperative symbiosis with fungi. (*b*) What is the name for a plant's response to gravity?

The objectives, text, and posttest items were reviewed by three colleagues of the authors in order to assure that each objective was clearly stated and that each posttest item was directly referenced to its appropriate objective. Minor revisions in wording resulted from this review.

Procedure

Subjects were randomly assigned to two treatment groups. The first group received one half of the objectives. These 12 objectives had previously been randomly selected from the full list of objectives. The second group received no objectives and were instructed to learn everything in the text.

The subjects were handed an instructional package containing general directions, the objectives (for the group receiving half of the objectives only), and the learning passage. The subjects had a maximum of 30 minutes in which to study the passage, although each subject controlled his own study time. During the learning task, the subjects were permitted to review any section of the text at their discretion. The subjects who had received objectives were further permitted to refer to them freely. When satisfied with his learning, each subject indicated in a space provided on his cover sheet the exact time, as indicated on the room clock. He then individually exchanged his instructional package for the posttest and additional, nonrelated reading material to keep him occupied until all subjects had completed the experiment. Posttest directions for the group receiving half of the objectives indicated to the subjects that they should try to answer all items, whether or not they were referenced to their objectives. Directions to the group receiving no objectives were simply to try to answer all items.

Results

The data collected in this study consisted of posttest scores and study latencies. The posttest scores were partitioned into two subscores: the first of these was referenced to the partial list of objectives received by the group receiving half of the objectives (relevant learning for that group) and the second subscore was referenced to the set of objectives not received by this group (incidental learning for that group).

The Kuder Richardson Formula 20 reliability indices for these subscores were .68 and .64. It is recognized that the use of instructional objectives and the implications thereof for a criterion-referenced approach should lead to the use of a criterion-based technique such as the one proposed by Livingston (1972). However, because there was only one test item per objective and because no percentage-type criterion was utilized, such a technique could not be employed.

Statistical contrasts using analysis of variance were made between the group receiving half of the objectives and the group receiving none of the objectives. These were made independently for each of the subscores. On Subscore 1, a significant effect for the availability of objectives was obtained ($F = 12.4$, $df = 1/56$, $p < .05$, $R^2 = 18\%$). Similar results resulted on Subscore 2 ($F = 23.3$, $df = 1/56$, $p < .05$, $R^2 = 29\%$). From Table 1, it can be seen that the group receiving half of the objectives performed better than the group receiving none of the objectives on Subscore 1, indicating that relevant learning was enhanced by the availability of objectives. Indeed, the difference of 2.3 points between the groups accounted for 18% of this subscore's variance. Incidental learning, on the other hand, was depressed, as evidenced by the means of 3.2 and 5.6, respectively, for the group receiving half of the objectives

TABLE 1
MEAN NUMBER CORRECT ON THE POSTTEST

No. objectives per group	Subscore 1[a]		Subscore 2[b]		Total score	
	M	*SD*	*M*	*SD*	*M*	*SD*
Half	7.4	2.7	3.2	1.9	10.6	3.2
None	5.1	2.1	5.6	1.9	10.8	3.3

[a] Relevant learning for the group receiving half of the objectives.

[b] Incidental learning for the group receiving half of the objectives.

and the group receiving none of the objectives. The difference of 2.4 between the groups accounted for 29% of Subscore 2 variance.

The difference between the two groups with respect to total score was not significant at the .05 level ($F < 1, df = 1/56$), nor was the difference between the time (in minutes) that each group spent studying the text ($F = 1.5, df = 1/56$). For the group receiving half of the objectives, the mean was 20 minutes ($SD = 4.8$), whereas it was 18.5 ($SD = 4.5$) for the group receiving none of the objectives. These figures represent the combined time involved in reading the directions, reading and referring back to the objectives (for the group receiving half of the objectives only), and studying the text.

Discussion

The results confirm the hypothesis elaborated for this research. The possibility that objectives have a focusing effect on learning seems to be supported by the fact that while the two groups did not appreciably differ either in total posttest score nor in study time, they did differ on each of the two subscores. The subjects who received half of the instructional objectives attained more of those objectives than their counterparts not provided with objectives. They furthermore attained fewer of the nonpresented objectives than their counterparts without objectives. It can be implied from these results that they used the objectives provided them in order to focus their learning on the relevant material (as perceived through their list of objectives) and to pay less attention to the incidental material (those parts of the material not referenced to their objectives). The subjects not provided with any objectives, on the other hand, engaged their learning equally on all parts of the material.

The results obtained in this study are in agreement with previous research (Morse & Tillman, 1972; Rothkopf & Kaplan, 1972) only with respect to relevant learning. In all three studies, objectives served to increase relevant learning. With respect to incidental learning, however, the present results conflicted sharply with the previous results. Morse and Tillman (1972) found no significant difference on incidental learning between a group with half of the objectives and a group without objectives. They concluded that objectives did not adversely affect incidental learning. Rothkopf and Kaplan (1972), on the other hand, found that objectives facilitated not only relevant learning but incidental learning as well.

As expressed in the introduction, these differences could stem directly from the fact that each of the three studies was actually dealing with a different population. In the Rothkopf and Kaplan (1972) report, no mention was made about the familiarity of the subjects with respect to the role of objectives in learning. It could be presumed that their subjects had little familiarity with learning objectives. Morse and Tillman (1972), on the other hand, trained a subset of their subjects on the use of objectives without apparent effect. In the present study, the participating subjects had practical experience in using objectives in one of their academic courses, an experience which the subjects in the other two studies presumably did not share. We could, therefore, be dealing with three distinct student populations: one having no familiarity with objectives and no experience with them, one familiar with objectives but lacking experience with them, and one with direct previous experience with objectives. The results obtained in each of the three studies may, therefore, be generalized only to their respective populations.

A further distinction between the studies was the type of learning which was involved in each of them. While the present study, as well as the Rothkopf and Kaplan (1972) study, used objectives subsumed mainly under the knowledge category of learning (Bloom, 1956), the Morse and Tillman (1972) objectives related to higher levels of learning (mainly Bloom's category entitled "Analysis").

Time

With respect to time, no hypothesis had been advanced in the present research. While the group receiving half of the objectives spent slightly more time studying the passage, the difference was not great (approximately 7% more time).

The correlations between time and performance were also low (about .25) and identical across groups. Carver (1970) strongly argued that the results obtained in orienting-stimuli research were generally confounded by time, since the subjects receiving orienting stimuli usually spent slightly more time studying the material. This criticism, however, does not have any implication for the results of the present study, since what was being investigated was not an overall effect on performance but rather a differential effect on performance. Both groups, in fact, performed equally well on the total score. Furthermore, with respect to time, it is questionable whether the laboratory studies dealing with orienting stimuli (including the present one) are representative of the situation involved in a regular academic setting. It could easily be expected that the effect of objectives on study time over an academic semester would be quite different from the effects obtained in laboratory studies of short duration.

Conclusion

The present research is seen as supportive of the hypothesis that objectives facilitate learning by both focusing the learning effort on relevant material and detracting attention from incidental material. The results obtained, however, are directly generalizable only to the knowledge category of learning and should be replicated with other types of learning.

Furthermore, the objectives utilized in this study were very specific objectives, which would be found in typical classroom situations only infrequently. It would be useful, therefore, to replicate these results in a setting more representative of the regular academic situation.

It should be noted also that the results were obtained in a situation in which objectives were developed from an existing text. One could expect different results in a situation in which objectives were developed first and then instructional materials were developed around the objectives. Indeed, much less incidental material would be present in such a case.

The present study once again points to the requirement that researchers in the field of instructional objectives insure that their subjects are familiar with objectives and actually use them if the results of their research are to be generalizable to an appropriate population. This factor was considered as one strong reason for not finding expected results in some of the previous research efforts dealing with objectives (Duchastel & Merrill, 1973).

References

Bloom, B. S. (Ed.) *Taxonomy of educational objectives: Cognitive domain.* New York: David McKay, 1956.

Carver, R. P. A critical review of mathemagenic behaviors and the effect of questions upon the retention of prose materials. Paper presented at the annual convention of the American Psychological Association, Miami Beach, September 1970.

Duchastel, P. C., & Merrill, P. F. The effects of behavioral objectives on learning: A review of empirical studies. *Review of Educational Research,* 1973, **43,** 53–69.

Frase, L. T. Some unpredicted effects of different questions upon learning from connected discourse. *Journal of Educational Psychology,* 1968, **59,** 197–201.

Frase, L. T., Patrick, E. M., & Schumer, H. Effect of question position and frequency on learning from text under different levels of incentive. *Journal of Educational Psychology,* 1970, **61,** 52–56.

Krieger, L. *The mushroom handbook* (2nd ed.) New York: Dover, 1967.

Livingston, S. A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement,* 1972, **9,** 13–27.

Morse, J. A., & Tillman, M. H. Effects on achievement of possession of behavioral objectives and training concerning their use. Paper presented at the American Educational Research Association annual convention, Chicago, April 1972.

Patrick, E. M. Prose learning: Induced question and response rehearsal, and question repetition. Unpublished manuscript, University of Massachusetts, 1968.

Rothkopf, E. Z. The concept of mathemagenic activities, *Review of Educational Research,* 1970, **40,** 325–336.

Rothkopf, E. Z., & Kaplan, R. An exploration of the effect of density and specificity of instructional objectives on learning from text. *Journal of Educational Psychology,* 1972, **63,** 295–302.

Tiemann, P. W. Student use of behaviorally-stated objectives to augment conventional and programmed revisions of televised college economics lectures. Paper read at the annual meeting of the American Educational Research Association, Chicago, April 1968.

(Received April 10, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 486-490

EFFECTS OF INSERTED QUESTIONS ON RETROACTIVE INHIBITION IN MEANINGFUL VERBAL LEARNING

BASIL S. WALKER[1]

University of New South Wales, Kensington, Australia

Subjects answered questions while reading the original-learning and interpolated-learning passages in the usual retroactive inhibition design. In Experiment 1, the inserted questions in the original-learning and related interpolated-learning passages were concerned with topics common to both passages, but in the case of the interpolated-learning passage, they indicated a competing response. In Experiment 2, the inserted questions for each passage covered material relevant only to that particular passage. In both cases, the presence of questions improved retention for material directly related to the questions but not for incidental information, and their presence had no effect on the amount of retroactive inhibition.

Recent studies in which the similarity relationship between the original-learning passage and the related interpolated-learning passage has been clearly specified and paralleled in the test items of the original-learning test have demonstrated that interference effects do operate in the retention of connected discourse (Anderson & Myrow, 1971; Crouse, 1971). Significant retroactive inhibition was obtained both when there was no measure of either original learning or interpolated learning prior to recall and when measures of both original learning and interpolated learning were made prior to recall (Crouse, 1971). Elimination of output had no effect on the amount of retroactive inhibition, but this comparison was cross-experimental and had different time intervals between original learning and recall. The presence of an immediate test after original learning but not after interpolated learning substantially improved delayed retention and sharply reduced interference effects (Anderson & Myrow, 1971). However, given that the level of original learning is increased by the presence of an immediate test even when no knowledge of results is provided (Roderick & Anderson, 1968), the reduction of interference is predictable from the relationship between the degree of original learning and that of interference (Slamecka & Ceraso, 1960) without any assumption of distinct test effects.

That the insertion of questions into both original-learning and related interpolated-learning passages might affect the amount of retroactive inhibition is suggested by the finding that the presence of test-like events in prose learning goes beyond direct-instructive effects and produces change in performance on reading material not directly related to the text components on which the test items were based (Rothkopf, 1970). If inserted questions are interpreted as having a review- rather than a "forward"-shaping function, then requiring subjects to answer questions during both original-learning and interpolated-learning passages may promote the process of "integrative reconciliation" (Ausubel, 1963) and thus reduce interference.

The purpose of the present study was to determine whether the amount of retroactive inhibition in prose learning would be affected by the presence of inserted questions in both original-learning and interpolated-learning passages.

[1] Requests for reprints should be sent to Basil S. Walker, Student Counseling and Research Unit, University of New South Wales, P.O. Box 1, Kensington N.S.W. 2033, Australia.

Experiment 1

Method

Passages and tests. The original-learning and related interpolated-learning passages, each describing a fictitious African tribe, were those used previously by Anderson and Myrow (1971).[2] Both passages covered the same topics, and parallel information was divided into three categories: same, unrelated, and conflicting. There were approximately equal amounts of the three types of information. The criterion test on the original-learning passage consisted of a 30-item, short-answer test and the same 30 items in a multiple-choice format. There were three types of test items corresponding to the three categories of information. Facilitation items contained an item stem which was represented in substantially the same form in both the original-learning and related interpolated-learning passages and which would be answered in the same way on the basis of either passage. Neutral items concerned material discussed only in the original-learning passage. Interference item stems were represented in both passages, but the responses associated with the items were different in the two passages. All of the interference items in the multiple-choice test included the specific competing response from the related interpolated-learning passage as one of the distractors. The multiple-choice test was corrected for guessing using the formula $R - W/3$, and the short-answer test was scored on the basis of a predetermined key which specified the limits of acceptable answers.

The unrelated interpolated-learning material was a passage on pestilence. The original-learning, related interpolated-learning, and unrelated interpolated-learning passages contained 2,240, 2,190, and 2,475 words, respectively.

Each learning passage was divided into three approximately equal sections. Each section within the original-learning passage and the related interpolated-learning passage contained roughly the same proportion of same, unrelated, and conflicting information. Questions, typed on a separate page and in short-answer form, were inserted after each section; four questions after Section 1 and three questions after Sections 2 and 3. The 10 interference items from the original-learning test served as inserted questions for the original-learning and related interpolated-learning passages, and 10 new questions were prepared for the unrelated interpolated-learning passage. Although the same questions appeared in the original-learning and related interpolated-learning passages, the set of questions after each section was different because similar topics were not covered in corresponding sections in the two passages.

Design. The design was a $2 \times 2 \times 2 \times 3$ analysis of variance with repeated measures on the last two factors. The between-subjects factors were presence or absence of inserted questions and type of interpolated-learning passage (related or unrelated). The within-subjects factors were response mode (short answer and multiple choice) and item type (facilitation, neutral, and interference).

Procedure. Both original- and interpolated-learning passages, with or without inserted questions, were placed in a single booklet and read in a single study session. Instructions preceding the passages told the subjects that they were to read carefully, rereading any difficult parts, but that they were not to turn back after having completed a page. They were informed that they would be required to answer questions at various points in the material or, in the absence of inserted questions, that they would be examined in the material at some later time. At the end of the last page of each section and at the bottom of each page of questions, the subject recorded, from a stop watch, the time taken to read that section or complete that page of questions. There was no time limit on reading or answering questions.

Seven days after their study session, the subjects took the short-answer and multiple-choice original-learning tests. Ten minutes was allowed for each form of the test.

Subjects. The subjects were 80 first-year undergraduate arts students (44 males and 36 females) who volunteered for participation in the study. Subjects were randomly assigned to the four experimental conditions with the restriction that the ratio of males to females in each condition be the same. The subjects were mostly run in small groups of from 6 to 10, with some subjects being done individually.

Results

Original learning. In terms of the time taken to read the original-learning passage, there was no significant difference for the main effects of inserted questions and type of interpolated-learning passage nor significant interaction between these effects (in all cases, $F < 1.00$, $df = 1/76$). Considering only the two groups who had inserted questions, there was no significant difference in the percentage of correct responses given to the inserted questions in the original-learning passage. The mean percentage correct for the unrelated interpolated-learning group was 93.0% and for the related interpolated-learning group, 93.5% ($F < 1.00$, $df = 1/38$).

Delayed recall. The mean percentages correct on parts of the delayed test for each experimental group appear in Table 1.

There was no overall effect of interpo-

[2] The author is indebted to R. C. Anderson for permission to use learning passages and tests.

TABLE 1

MEAN PERCENTAGE CORRECT ON THE DELAYED ORIGINAL-LEARNING TEST

Group	n	Item					
		Short answer			Multiple choice		
		Facilitation	Neutral	Interference	Facilitation	Neutral	Interference
Related interpolated learning							
No inserted questions	20	46.0	45.0	39.0	60.7	62.0	47.8
Inserted questions	20	43.5	48.0	67.5	63.0	60.5	66.7
Unrelated interpolated learning							
No inserted questions	20	35.0	52.5	52.0	53.3	62.2	67.8
Inserted questions	20	42.0	47.5	76.5	55.7	60.2	84.7

lated learning, but the Interpolated-Learning Passage × Item Type interaction was significant ($F = 16.81$, $df = 2/152$, $p < .01$). There was an overall effect of inserted questions ($F = 7.83$, $df = 1/76$, $p < .01$), but the Interpolated-Learning Passage × Item Type × Inserted Questions interaction, represented in Figure 1, was not significant. The interaction of interpolated-learning passage and item type conforms to the interference model; the expected retroactive inhibition on the response-different items and facilitation on the response-same items was obtained. The presence of inserted questions increased retention of interference items by nearly the same amount in the related interpolated-learning and unrelated interpolated-learning conditions and, except for small nonsignificant changes in the unrelated interpolated-learning condition (see Figure 1), had no effects on facilitation or neutral items.

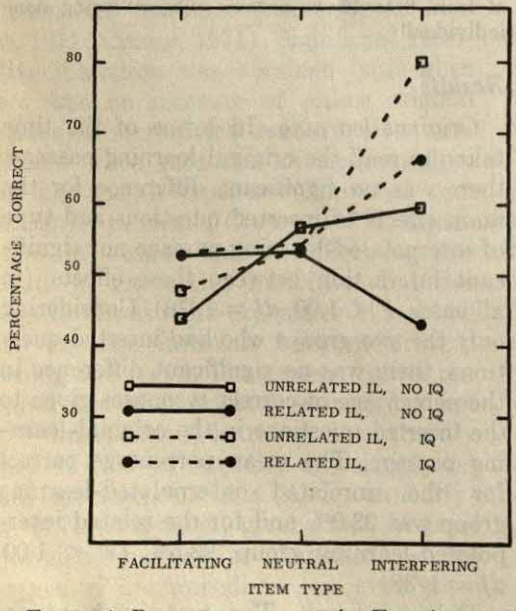


FIGURE 1. Percentage correct in Experiment 1 on the delayed original-learning test as a function of the interpolated-learning passage (IL), the presence or absence of inserted questions (IQ), and the item type.

Neither the Interpolated-Learning Passage × Response Mode × Item Type interaction nor the Inserted Questions × Response Mode × Item Type interaction was significant, but retroactive inhibition was much more marked on the multiple-choice than on the short-answer portion of the delayed test.

EXPERIMENT 2

Since in Experiment 1 the interference items from the original-learning test served as inserted questions in the original-learning and related interpolated-learning passages, any nondirect-instructive effects of the inserted questions on retroactive inhibition was confounded with direct-instructive effects. In Experiment 2, the questions inserted into the original-learning, related interpolated-learning, or unrelated interpolated-learning passages were concerned with information present only in that particular passage.

Method

Instead of the 10 interference items from the original-learning test, the 10 neutral items were

inserted into the original-learning passage. Ten new questions, concerned with information contained in the related interpolated-learning passage only, were prepared and inserted into that passage. In all other respects, the materials and designs were the same as in Experiment 1.

Subjects and procedure. The subjects were male, high school students who were in Grade 11. Seventy-four subjects were present for both the study session and the delayed-recall test. Subjects within each of the two Grade 11 classes involved in the study were randomly assigned to one of the four experimental conditions by passing out material stacked in a random order. Except for the fact that the time taken to read each section of each passage and to answer the inserted questions was not recorded, the rest of the procedure was the same as in Experiment 1.

Results

The significant effects were the same as in Experiment 1. The related interpolated-learning passage produced retroactive inhibition ($F = 26.37$, $df = 2/140$, $p < .01$) for the Interpolated-Learning Passage × Item Type interaction but, again, the Interpolated-Learning Passage × Inserted Questions Item Type interaction, shown in Figure 2, was not significant.

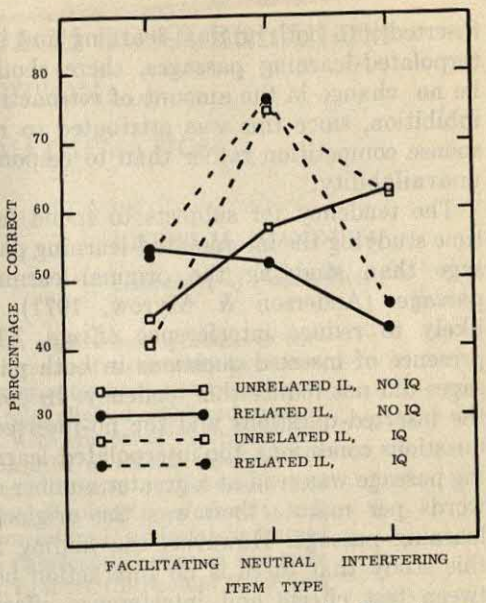


FIGURE 2. Percentage correct in Experiment 2 on the delayed original-learning test as a function of the interpolated-learning passage (IL), the presence or absence of inserted questions (IQ), and the item type.

Discussion

Although the presence of inserted questions improved delayed recall, they had no affect on the amount of retroactive inhibition, and the improvement in delayed recall was restricted to the material on which the inserted questions were based.

In further confirmation of earlier studies (Anderson & Myrow, 1971; Myrow & Anderson, 1972), retroactive inhibition was attributable to response competition rather than to response unavailability. Rather than retroactive inhibition being greater on the short-answer than on the multiple-choice test, as required by the response availability hypothesis, it was greater on the multiple-choice test, although not significantly so. Moreover, analysis of errors on the multiple-choice interference items showed that for the related interpolated-learning groups, 60.0% of the errors in the no-inserted-questions conditions and 68.0% of the errors in the inserted-questions condition resulted from choosing the specific competing interpolated-learning passage distractor. The corresponding figures for the unrelated interpolated-learning groups were 21.7% and 17.8%.

The improvement in performance in delayed recall as a result of the presence of inserted questions is attributed to the strengthening of specific associative connections. In the Anderson and Myrow (1971) study, it was concluded that completing an immediate test primarily influences response availability rather than specific associative connections because the immediate test differentially affected performance on short-answer and multiple-choice tests. However, in the Anderson and Myrow study, the immediate test included multiple-choice items, whereas in the present study the inserted questions were in a short-answer format and thus could only affect items for which responses were already available.

The increase in performance due to the presence of short-answer inserted questions was substantial, but it is possible that further improvement could have been obtained by including multiple-choice items. However, if the multiple-choice questions were

inserted into both original-learning and interpolated-learning passages, there should be no change in the amount of retroactive inhibition, since this was attributed to response competition rather than to response unavailability.

The tendency for subjects to spend less time studying the interpolated-learning passage than studying the original-learning passage (Anderson & Myrow, 1971) is likely to reduce interference effects. The presence of inserted questions in both passages did not reduce this tendency. In both the inserted-questions and the no-inserted-questions conditions, the interpolated-learning passage was read at a greater number of words per minute then was the original-learning passage. However, the finding in this study that there is no interaction between test effects and interference effects means that tests can be used to determine the extent to which the level of learning in the original-learning and interpolated-learning passages differ, with no fear of confounding test effects and interference effects.

REFERENCES

Anderson, R. C., & Myrow, D. L. Retroactive inhibition of meaningful discourse. *Journal of Educational Psychology*, 1971, **62**, 81–94.

Ausubel, D. P. *The psychology of meaningful verbal learning*. New York: Grune & Stratton, 1963.

Crouse, J. H. Retroactive interference in reading prose materials. *Journal of Educational Psychology*, 1971, **62**, 39–44.

Myrow, D. L., & Anderson, R. C. Retroactive inhibition of prose as a function of the type of test. *Journal of Educational Psychology*, 1972, **63**, 303–308.

Roderick, M. C., & Anderson, R. C. Programmed introduction to psychology versus textbook-style summary of the same lesson. *Journal of Educational Psychology*, 1968, **59**, 381–387.

Rothkopf, E. Z. The concept of mathemagenic activities. *Review of Educational Research*, 1970, **40**, 325–336.

Slamecka, N. J., & Ceraso, J. Retroactive and proactive inhibition of verbal learning. *Psychological Bulletin*, 1960, **57**, 449–475.

(Received April 17, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 491-498

THE ALGORITHMIC APPROACH TO CURRICULUM CONSTRUCTION: A FIELD TEST IN MATHEMATICS[1]

WALTER EHRENPREIS
Trenton State College

JOSEPH M. SCANDURA[2]
University of Pennsylvania

This study constitutes a field test of an algorithmic approach to curriculum construction. First the method was used successfully to characterize the content inherent in a mathematics textbook in terms of rules (algorithms). The use of higher-order rules allowed a 40% reduction in total rules. Two rule-based curricula were compared experimentally. The discrete (D) rules curriculum consisted of all 303 (lower-order) rules. The higher-order (H) curriculum included 169 of these rules plus 5 higher-order rules. The Curriculum H subjects performed as well with the eliminated rules as did the Curriculum D subjects who learned the rules directly, and they performed significantly better on new tasks beyond the scope of either curriculum. In short, Curriculum H subjects were taught less but learned more.

Curriculum construction has traditionally been an artistic endeavor. Even today, the vast majority of texts and new curricula are developed almost exclusively on the basis of the curriculum constructor's subject-matter knowledge and professional know-how.

During the 1960s, a strong technological counterforce developed under the leadership of behavioral scientists. The basic position taken was that objectives must be stated in behavioral (operational) terms so as to make it possible to determine, through testing, whether learners have achieved individual objectives. As a result, a healthy debate developed between proponents of behavioral objectives (e.g., Gagné, 1970; Lipson, 1967; Mager, 1962; Popham, 1969; Tyler, 1964) and others who raised cautions concerning their use (e.g., Atkin, 1968; Ebel, 1970; Eisner, 1967).

In recent publications, the positions taken have become increasingly more flexible (e.g., Glaser, 1973; Resnick, 1972; MacDonald-Ross, 1973). A statement by Scandura (1971) summarizes much of the current view: "It is felt that complete reliance on operationally defined objectives has led some to fragmented curricula, curricula based on discrete bits of knowledge [p. 4]."

Elaborating on this view, Scandura (1972) has identified two basic inadequacies of the behavioral objectives approach considered in its simplest form: (*a*) The approach deals only with observable behavior and says little about how that behavior is to be generated, and (*b*) it provides no systematic way of dealing with interrelationships among the identified objectives, or equivalently, of building transfer into a curriculum.

Regarding Point *a*, the distinction between the behavior of a subject and the knowledge (rule) that makes that behavior possible is fundamental. It can easily be proven mathematically that if just one rule exists for generating a class of behaviors, then there is an infinite number of other rules that will do the same thing (e.g., see Rogers, 1967). This fact is important in

[1] This article is based on a PhD dissertation in 1972 by the first author under the second author's chairmanship at the University of Pennsylvania. This study was supported, in part, by National Science Foundation Grant GW-6796 to the second author.

[2] Requests for reprints should be sent to Joseph M. Scandura, Graduate School of Education, University of Pennsylvania, Philadelphia, Pennsylvania 19174.

curriculum planning because in practice there is almost always more than one viable way of approaching a task. The subtraction methods of borrowing and equal additions, for example, are both widely used.

Regarding Point *b*, it is clearly an impossible task, with any but the most trivial curricula, to explicitly teach the learner all that the curriculum constructor wants him to know. The limitations imposed by time and the capacity of the learner to absorb and retain information make this impractical. Some attention to interrelationships would seem almost essential.

One approach to this problem is based on learning hierarchies (e.g., Gagné, 1970; Resnick, Wang, & Kaplan, 1970). As is well known, this approach makes use of task analysis (e.g., Miller, 1962) as a means of determining subordinate tasks. Subordinate tasks are prerequisite to so-called Higher Order 1 tasks in the sense that transfer to Higher Order 1 tasks frequently occurs once all of the prerequisites are learned. [It may be noted parenthetically that implicit in any specific task analysis is a specific underlying rule (which is not necessarily, and often is not, explicit in the analyst's mind). Since different rules may underlie the same task (as indicated above) it follows that any given task, theoretically speaking, may be task analyzed in any number of different ways (e.g., consider task analyses of subtraction based on borrowing and equal additions).]

In the present study, we have adopted a second (algorithmic) approach to the problem of transfer (Scandura, 1972). This approach bears some relationship to learning hierarchies, but it includes an important conceptual generalization that has often gone undetected because of the common use of the descriptor "higher order."

This approach is based on a recent theory of structural learning (Scandura, 1973) in which Higher Order 2 rules may operate on other rules (e.g., subordinate ones) to generate what in task analysis are Higher Order 1 rules. [Mathematically speaking, the distinction is precisely that between a function (Higher Order 2) and a function value or output (Higher Order 1).] Unlike the "directions," which are sometimes necessary in moving from one level in a hierarchy to another, Higher Order 2 rules operate on *classes* of rules and are not limited to rules in particular hierarchies. Higher Order 2 rules may operate between various levels of any one of a class of hierarchies, including hierarchies that appear superficially quite different. Consider, for example, the following two, simple Higher Order 1 tasks: (*a*) Given a certain number of yards, find the equivalent number of inches; and (*b*) Given an airplane (on the ground), get it up in the air and then back down (safely). One way to analyze these tasks is to break each task into component parts: (*a*) one subordinate task for converting yards into feet and another for feet into inches and (*b*) subordinate tasks for "taking off" and "landing." Although they involve quite different rules, the hierarchies have a common structure and, in particular, the Higher Order 1 tasks (more exactly, the rules for solving the indicated Higher Order 1 tasks) can be generated from the respective subordinate task (rules) by applying a Higher Order 2 *composition* rule (for details, see Scandura, 1973b, pp. 213–218). This Higher Order 2 composition rule would be reflected in typical learning hierarchies by parallel but different "instructions." In effect, a single, Higher Order 2 rule can take the place of a potentially infinite class of separate "instructions."

Further, Higher Order 2 rules are not limited in their application to traditional hierarchies. For example, given a rule for converting inches to centimeters (1 inch = 2.54 centimeters), it is a simple matter to envision a Higher Order 2 *inverse* rule which applies to such rules and generates their inverses. In this case, the output would be a rule for converting centimeters to inches (1 centimeter = inch/2.54). For a discussion of these and other differences between Higher Order 1 and Higher Order 2 rules, see Scandura (1973a). In this article, all subsequent references to "higher order" are of Type 2.

The algorithmic approach to curriculum construction is based on these ideas and dis-

tinctions. Specifically, the approach builds on the notion that each behavioral objective corresponds directly to a class of tasks that can be computed (solved) by applying a rule or algorithm. It is further assumed that curricula (i.e., what is to be learned) can be represented in terms of finite sets of rules, including higher-order rules that operate on rules. In effect, the task of curriculum construction may be viewed as one of identifying a finite set of rules that provides an efficient account of the desired behaviors. The algorithmic approach is basically a method for devising curricula based on behavioral objectives and characterized in terms of rules and higher-order rules.

The first step in this method is to select text materials to analyze. This, of course, involves making value judgments concerning the type of material to be considered. All of the tasks implicit in the text material are then identified and stated as behavioral objectives. Next, rules are written for solving each of the tasks, and parallels among these rules are identified. Such parallels are indicative of common structure and provide a basis for devising higher-order rules. Finally, those rules that are derivable by application of the higher-order rules to other rules in the characterizing set may be eliminated.

To summarize, the algorithmic approach provides a potential basis for overcoming the two, aforementioned major *limitations* of the simple behavioral objectives approach. First, it makes specific what the subject must learn in order to demonstrate mastery on behavioral objectives and, thus, might provide a viable basis for instruction. Second, it makes explicit provision for the inclusion of higher-order relationships among objectives. Indeed, it provides a systematic way for possibly building transfer potential into a curriculum.

Two studies are reported. The first study was strictly analytical in nature and was designed to determine the general feasibility of the algorithmic approach. Specifically, we wanted to determine the practicality of characterizing the knowledge inherent in a given mathematics text in terms of a finite set of rules, including higher-order rules.

The second study was contingent upon the success of Study 1. The purpose of this study was to determine (*a*) whether making rules explicit provides a viable basis for instruction in the classroom, and (*b*) whether the introduction of higher-order rules provides an adequate basis for improving the ability of students to transfer.

Study 1

Feasibility of the Algorithmic Approach to Curriculum Construction

In order to judge the feasibility of the approach, the following two criteria were established:

1. A subjective appraisal of (*a*) the ease with which the tasks (behavioral objectives) inherent in the given text material could be identified, (*b*) the ease with which rules associated with each of the respective tasks could be written, and (*c*) the extent to which the tasks and rules identified were compatible with the approach taken in the text.

2. The extent and ease with which the higher-order rules inherent in the text could be (*a*) identified and (*b*) used to eliminate those rules (and their corresponding tasks) that were derivable by application of the higher-order rules to other identified rules.

Consideration was also given to the sheer numbers of tasks and rules involved and particularly to the extent these numbers could be reduced by the introduction of higher-order rules.

Method

Part 3 of *Mathematics: Concrete Behavioral Foundations* (Scandura, (1971) was chosen for analysis. (This book was later analyzed in its entirety and was published as a workbook, Scandura, Durnin, Ehrenpreis, Luger, 1971.)

The first step was to identify the individual tasks (behavioral objectives) inherent in the text. This was accomplished by going through the text paragraph by paragraph and asking what performance capabilities might reasonably be expected of a student who had studied the material. For example, the following tasks were identified in *Mathematics: Concrete Behavioral Foundations* on pages 182 and 191, respectively.

Task A. Given a whole number m, determine whether or not x is an additive identity for m.

Task B. Given a whole number m, determine whether or not y is a multiplicative identity for m.

The second step was to identify and eliminate redundancies in the tasks identified. Very few (less than 8) such redundancies were found in the text aside from prerequisites to other tasks. Prerequisite tasks were not eliminated because it was felt desirable to maintain the original sequencing of ideas in the text.

Third, one efficient rule was constructed for each task. The rules were written so as to be compatible with the text materials. The rules written for the illustrative Tasks A and B were as follows:

Rule A. Find the sum $m + x$ and then the sum $x + m$. If $m + x = x + m = m$, then x is an additive identity for m; if $m + x \neq m$ or $x + m \neq m$, then x is not an additive identity for m.

Rule B. Find the product $m \times y$ and then the product $y \times m$. If $m \times y = y \times m = m$, then y is a multiplicative identity for m; if $m \times y \neq m$ or $y \times m \neq m$, then y is not a multiplicative identity for m.

The fourth step was to look for higher-order relationships among the rules. These relationships were stated as tasks to be performed, and higher-order rules underlying these (higher order) tasks were then constructed. For example, Rules A and B are obviously related. The nature of this relationship can be illustrated by the following task and its underlying higher-order rule.

Task H. Give a rule for demonstrating that a given set of numbers provides an instance of a property (e.g., commutativity) under some operation, generate a corresponding rule involving another operation.

Rule H. In the given rule, replace the original operation by the new operation and any "special" element (e.g., the identity) by its counterpart.

Fifth, the higher-order rules identified in Step 4 were used to eliminate those tasks and corresponding rules that are derivable by application of the higher-order rules to other rules identified. Thus, for example, Rule B was eliminated inasmuch as it could be generated by applying Rule H to Rule A.

Results and Discussion

The chapters analyzed lent themselves very naturally to a task–rule type of analysis. The major requirements found necessary for such analysis were a thorough familiarity with the subject matter and a good working knowledge of the algorithmic approach.

Upon completion of Step 3, a list of 303 tasks and their corresponding rules (one rule for each task) had been identified. A list containing 174 rules was obtained upon completion of Step 5. Five of these were higher-order rules acting on rules concerned with (*a*) ordered sets (ordinal numbers) and unordered sets (cardinal numbers), (*b*) well-defined operations, (*c*) properties of number systems (e.g., commutativity), (*d*) inverse operations, and (*e*) particularization (i.e., assigning values to variables). The first higher-order rule made it possible to eliminate 4 rules; the second, 17; the third, 84; the fourth, 7; and the fifth, 22.

Although the process of identifying the "tasks" inherent in the text was time-consuming, it was felt that the list of 303 tasks gave almost complete coverage of the material. Once the tasks were identified, there was little difficulty encountered in writing rules compatible with the approach taken in the text. This was partly due to the fact that the narrative and illustrative examples provided adequate guidance. The identification of the higher-order rules and the subsequent reduction of the list of 303 rules to the final list of 174 rules was the most difficult step. Some of the higher-order rules (e.g., ordered sets and unordered sets) were easier to identify than others (e.g., particularization), and the analysis was pursued just far enough to demonstrate the feasibility of the approach. Doubtless, a more intensive analysis would have resulted in a larger number and variety of higher-order rules. Overall, based on the criteria established, it was concluded that the algorithmic approach was feasible.

Study 2

Evaluation of the Algorithmic Curriculum

The second study, in contrast to the first, was experimental and involved a comparison of two rule-based curricula. The first curriculum (D) was characterized in terms of a list of *discrete* tasks and rules for solving these tasks, one rule for each task. The second curriculum (H) included the *higher-*

order rules and all Curriculum D tasks and rules except those derivable by application of the higher-order rules to other rules in Curriculum D (as described in Step 5, Study 1).

The degree of learning evidenced by students trained in Curriculum D provided evidence concerning the hypothesis that making rules explicit provides a viable basis for classroom instruction. This hypothesis was tested directly in terms of mastery rather than by comparison with some arbitrarily defined control.

Comparative performance of the students trained in Curricula D and H pertained to the question of transfer. We hypothesized that students trained in the higher-order rules (Curriculum H) would perform on the Curriculum D tasks that had been *eliminated* from Curriculum H as well as the subjects who had been trained on the Curriculum D tasks. This was expected (according to the theory) because these tasks could be solved by using higher-order rules to derive solution rules for these tasks. Furthermore, we predicted that the Curriculum H students would perform significantly better on tasks not in Curriculum D that could be solved in the same way (i.e., by use of the higher-order rules). No difference in performance was expected on tasks included in both curricula.

Method

Materials. The experimental materials were based directly on the analyses reported in Study 1. The discrete rules curriculum (D) resulted upon application of Steps 1–3 of the algorithmic approach. The higher-order rules curriculum (H) included Steps 4–5 as well. Curriculum D consisted of 303 tasks and rules, and Curriculum H of 174 task and rules including 5 higher-order tasks and rules. Although experimental materials were prepared for Chapters 5–9 of *Mathematics: Concrete Behavioral Foundations,* only those materials pertaining to Chapters 5, 6, and 7 were actually used in the experiment.

The two curricula were reproduced in the form of workbooks with the following format: First, a statement of the task. Second, a rule statement in simple terms. Next, from three to five worked examples (depending on the experimenter's judgment as to task difficulty), and finally, 10 exercises. In addition, a set of task exercises for review purposes was selected on the basis of the apparent difficulty subjects encountered during learning. All higher-order tasks were included in the Curriculum H review.

Pre- and posttests were also constructed. The pretest consisted of a sample of 40 exercises (stated exactly as they appeared in the workbooks) which (*a*) tested applications of rules found in *both* treatments, and (*b*) in the judgment of the experimenter, were most likely to discriminate among subjects. The posttest consisted of a stratified random sample of 32 task exercises from each of the following categories: (*a*) tasks found in Curricula D and H, (*b*) tasks found only in Curriculum D, (*c*) higher-order tasks found only in Curriculum H, and (*d*) tasks that were found in neither treatment but which theoretically could be *derived* from rules found in Curriculum H. The exercises were selected randomly from each of these categories—10 each from Categories (*a*) and (*b*), and 6 each from Categories (*c*) and (*d*). (Kuder-Richardson Formula 20 was used to obtain reliability coefficients of .86 for both the pretest and the posttest. High curricular (content) validity (Cureton, 1951) was assured by the method of construction used.)

Subjects and design. The subjects were 48 (16 male, 32 female) Trenton State College summer school students enrolled in two sections of a course on teaching modern mathematics in the elementary grades. The first author was the instructor for both sections of the course.

Pretest scores were used to assign 24 subjects to a high group and 24 subjects to a low group. Within each group, 12 subjects were randomly assigned to each curriculum.

Procedure. During the first class meeting of each section, the subjects were told that the course was experimental and that, with the exception of unrelated outside reading, all of their work would be done individually and during class time. Then, the pretest was administered. The subjects had all the time they needed to complete the test. The highest score on the pretest was 24 exercises correct out of the possible 40. The median was 10 and the mean was 11.2. Forty-five of the 48 subjects had pretest scores below 20, and 26 had scores of 10 or below.

Each subject purchased a notebook at the beginning of the six-week term; all experimental work was done in these notebooks. The instructional workbooks and student notebooks were distributed at the beginning of each two-hour, five-minute class

TABLE 1

NUMBER OF TASKS IN CHAPTERS 5 THROUGH 9

Curriculum	Chapter				
	5	6	7	8	9
Discrete rules	47	22	79	72	83
Higher order	32	17	46	44	35

period and collected at the end. The Curriculum D and Curriculum H subjects met in separate classrooms for three class periods per week. Each subject was permitted to go on to the next task as soon as he had demonstrated success on a number of consecutive exercises equal to the number of illustrative examples constructed for that task. A daily record of each subject's performance was kept, as were anecdotal records indicating particular difficulties or problems that arose.

During the first meeting of each week (beginning with Week 2), subjects were given a set of individualized review exercises; the exercises selected were those exercises completed during the preceding two weeks which had given the subject the greatest difficulty. The subjects were required to work these exercises without the use of the workbook or their notebooks. On the final class meeting of each week, the review exercises and solutions were returned to the subjects. They were told to check their solutions using their workbooks and notebooks and to correct any errors. As soon as this had been completed, the review materials were collected.

During the final three weeks, class time was set aside to enable subjects to review and study the rules they had learned up to that time (45 minutes during Weeks 4 and 5; 3 hours during Week 6). The amount of time spent working on the tasks and on review was the same for all subjects. In those few cases where a subject missed a class, he was required to make up the time missed.

At the end of the term, 14 of the 24 subjects in Curriculum D were working in Chapter 7, with the farthest advanced to Task 36. Two of the 24 Curriculum H subjects completed Chapter 7, and 16 others were working in Chapter 7. All Curriculum H subjects completed the five higher-order tasks.

The posttest was given during the next-to-last class meeting. The subjects were given all of the time they needed and they were encouraged to do their best.

Results and Discussion

Mastery. The Curriculum D subjects were successful on 339 of the 362 posttest tasks on which they as individuals had been trained, for a mastery level of 94%. (The number of test items on which each subject had been trained was determined by examination of his class workbook.) The Curriculum H subjects were successful on 183 of the 190 lower-order tasks (96%) on which they had been trained. It would appear that making rules explicit does provide a viable basis for instruction.

On the higher tasks, Curriculum H subjects performed as expected at a significantly higher level ($F = 7.37$, $df = 1,44$, $p < .01$) than the Curriculum D subjects. The overall means were 3.2 (Curriculum D) and 4.3 (Curriculum H) with a maximum score of 6. The data of the individual subjects showed that the Curriculum H subjects were successful on 103 out of the 144 posttest exercises (71.5%), whereas the Curriculum D subjects were successful on 77 out of 144 (53.4%) exercises. These proportions differed significantly (arc sine transformation, $Z = 2.9$, $p < .01$; cf. Baggaley, 1957).

It is of interest, nonetheless, that the Curriculum D subjects, although not trained on the higher-order tasks, did perform successfully 53.4% of the time. This suggests that some of the Curriculum D subjects may have known these (relatively simple) higher-order rules prior to the experiment, while others may have been able to induce them as they worked through Curriculum D. In any case, the performance level did not approach that of the Curriculum H subjects, and it can safely be concluded that training on higher-order rules had a positive effect. This gap undoubtedly would have been even greater had more sophisticated higher-order rules been introduced.

Transfer. The hypotheses pertaining to transfer were equally conclusive. As expected, Curriculum H subjects performed just about as well on tasks found only in Curriculum D as did the Curriculum D subjects who were trained on these tasks directly. The Curriculum D subjects were successful on 168 out of the 181 posttest tasks (93%) in this category, and the Curriculum H subjects were successful on 167 out of the corresponding 190 tasks (88%). The difference between these two proportions was not significant (arc sine transformation, $0 < Z < .5$, $p > .05$). The overall means were 8.3 (Curriculum D) and 8.0 (Curriculum H) with a maximum score of 10 ($F = 1.02$, $df = 1,44$, $p > .05$).

In addition, the Curriculum H subjects, as predicted, performed at a significantly higher level ($F = 30.03$, $df = 1,44$, $p < .001$) than the Curriculum D subjects on tasks beyond the scope of either curriculum. On the six tasks where solution rules could be derived from given rules via the higher-

order rules, the obtained means were 3.1 (Curriculum D) and 4.1 (Curriculum H). The Curriculum H subjects were successful on 98 of the 144 transfer opportunities provided (68%), and the Curriculum D subjects were successful on 74 of the 144 (51.3%). These proportions differed significantly (arc sine transformation, $Z = 2.9$, $p < .01$).

Relation between transfer and mastery on higher- and lower-order rules. It is also worth noting that the difference in group performance of about 17%–18% on the higher-order tasks was directly reflected in performance on the six transfer tasks that were neither in Curriculum D nor Curriculum H. There too the difference was about 17%–18%. This observation suggests that subjects who had (directly or indirectly) learned a higher-order rule were able to apply it successfully to transfer tasks. In effect, it would appear that the availability of a suitable higher-order rule, together with appropriate lower-order rules, provides a sufficient basis for transfer to new tasks.

A more detailed analysis showed that (*a*) of the 235 cases where subjects were successful on a higher order task, they were successful on 166 of the corresponding transfer tasks, giving 71% correct prediction; and (*b*) of the 101 cases where subjects were unsuccessful on a higher-order task, they were unsuccessful on 72 of the corresponding transfer tasks, again giving 71% correct prediction. These findings are generally compatible with a number of related "laboratory" experiments (Scandura, 1973b). Although predictability did not approach the levels obtained there (86%–100% accuracy), these results demonstrate the robustness of the theory.

Other data showed that transfer was not affected by the direct presentation on the transfer tasks of the necessary lower-order rules. (The rules to which the higher-order rules applied were presented directly on four of the six transfer tasks.) The Curriculum D subjects were successful on 48 out of 96 transfer problems (50%) in which the needed lower-order rule was presented and 28 out of 48 transfer problems (54%) in which the subjects were trained on the needed lower-order rule but were not formally presented with it on the test. Correspondingly, the Curriculum H subjects were successful on 66 out of 96 transfer problems (69%) and 32 out of 48 transfer problems (67%). Neither pair of proportions differed significantly ($Z = .59$, $p > .05$ for the Curriculum D differences and $Z = .29$, $p > .05$ for the Curriculum H differences). This suggests that memory was not an essential factor, at least under the present conditions where the level of mastery on the lower-order rules was about 95%.

Summary and Implications

In summary, these results clearly show that rules provide a viable and explicit basis for instruction and transfer. The Curriculum H subjects not only had fewer rules to learn than did the Curriculum D subjects, but they were also able to solve tasks that the other subjects could not. The strong transfer effects were obtained even though the six-week course came to an end just as many of the Curriculum H subjects were reaching that portion of the workbook where the higher-order formulation had made it possible to eliminate large numbers of Curriculum D tasks (cf. Table 1). If time had permitted some of the subjects to complete Chapters 8 and 9, it seems quite possible that an even greater difference might have been obtained in favor of the Curriculum H subjects in amount of material covered per unit of time.

These results suggest that the algorithmic approach should be given serious consideration in planning future curriculum development. Curricula that are characterized in terms of rules and higher-order rules provide an explicit basis for instruction and, even more important, make specific provision for remote transfer, something which many subject-matter specialists feel is lacking in current curricula based on operational objectives.

The present study, of course, *demonstrates* the viability of the algorithmic approach *only* with respect to certain parts of mathematics. Further research is needed to determine the feasibility of the approach with other subject matters. A study on criti-

cal reading by Lowerre and Scandura (1973) is particularly relevant in that it was based on a short-cut approximation of the algorithmic approach to curriculum development called dimensional analysis.

It also should be emphasized that while knowledge can be rigorously characterized in terms of rules, this does not imply that knowledge must be imparted to students (e.g., young children) in this manner. As in this study, rules can be acquired by telling or by discovery from instances or, in other cases, by symbol juggling or concrete manipulation. The choice is up to the teacher and depends on factors other than the particular knowledge in question. Instructional formats other than workbooks, of course, should also be explored. The important point is that if we know precisely what (rule) it is that we want a child to learn, then we can facilitate learning far better than if we do not.

REFERENCES

Atkin, M. J. Behavioral objectives in curriculum design: A cautionary note. *The Science Teacher,* 1968, **35,** 27–30.

Baggaley, A. R. A table to facilitate comparison of proportions by slide rule. *Psychological Bulletin,* 1957, **54,** 250–252.

Cureton, E. E. Validity. In E. F. Lindquist (Ed.), *Educational Measurements.* Washington, D.C.: American Council on Education, 1951.

Durnin, J., & Scandura, J. M. An algorithmic approach to assessing behavior potential: Comparison with item forms and hierarchical analysis. *Journal of Educational Psychology,* 1973, **65,** 262–272.

Ebel, R. L. Behavioral objectives: A close look. *Phi Delta Kappan,* 1970, **52,** 171–173.

Eisner, E. W. Educational objectives: Help or hindrance? *School Review,* 1967, **75,** 250–266.

Gagné, R. M. *The conditions of learning.* (2nd ed.) New York: Harper & Row, 1970.

Glaser, R. Educational psychology and education. *American Psychologist,* 1973, **28,** 557–566.

Lipson, J. I. Individualized instruction in elementary mathematics. In J. M. Scandura (Ed.), *Research in mathematics education.* Washington, D.C.: National Council of Teachers of Mathematics, 1967.

Lowerre, G., & Scandura, J. M. Conceptually based development of individualized materials for critical reading based on logical inference. *Reading Research Quarterly,* 1973–1974, **9,** 186–205.

MacDonald-Ross, M. Behavioral objectives—A critical review. *Instructional Science,* 1973, **2,** 1–52.

Mager, R. F. *Preparing instructional objectives.* Palo Alto, Calif.: Fearon, 1962.

Miller, R. B. Task description and analysis. In R. M. Gagne (Ed.), *Psychological principles in system development.* New York: Holt, Rinehart & Winston, 1962.

Popham, W. J. Probing the validity of arguments against behavioral goals. In G. W. Faust, M. C. Roderick, D. J. Cunningham, & R. C. Anderson (Eds.), *Current research on instruction.* Englewood Cliffs, N.J. Prentice-Hall, 1969.

Resnick, L. B. Open education: Some tasks for technology. *Educational Technology,* 1972, **12,** 70–76.

Resnick, L. B., Wang, M. C., & Kaplan, J. Behavior analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. Pittsburgh, Penn.: University of Pittsburgh, Learning Research and Development Center, 1970. (Monograph 2)

Rogers, H. J. *Theory of recursive functions and effective computability.* New York: McGraw-Hill, 1967.

Scandura, J. M. *Mathematics: Concrete behavioral foundations.* New York: Harper & Row, 1971.

Scandura, J. M. Plan for the development of a conceptually based curriculum for disadvantaged youth, Parts I and II. *Instructional Science,* 1972, **1,** 247–263, 363–387.

Scandura, J. M. On higher order rules. *Educational Psychologist,* 1973, **10,** 159–160. (a)

Scandura, J. M. *Structural Learning I: The theory and empirical research.* New York: Gordon & Breach, 1973. (b)

Scandura, J. M., Durnin, J., Ehrenpreis, W., & Luger, G. *Algorithmic approach to mathematics: Concrete behavioral foundations.* New York: Harper & Row, 1971.

Tyler, Ralph W. Some persistent questions on the defining of objectives. In C. M. Lindvall (Ed.), *Defining educational objectives.* Pittsburgh, Penn. University of Pittsburgh Press, 1964.

(Received June 11, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 499-511

INTERNALIZATION OF FILMIC SCHEMATIC OPERATIONS IN INTERACTION WITH LEARNERS' APTITUDES[1]

GAVRIEL SALOMON[2]

The Hebrew University of Jerusalem, Israel

Two major hypotheses were tested in three experiments. The first hypothesis was that youngsters can imitate and internalize filmic codes to be used subsequently as covert schematized mediators. The second hypothesis was that learners with low, relevant aptitude scores profit more from films which model for them schematic operations to be internalized than do high-aptitude learners. Two kinds of operations were either modeled, short-circuited, or not shown at all, thus requiring subjects to activate them on their own. These operations were zooming in on details (Experiments 1 and 2) and laying out of solid objects (Experiment 3). Subjects were 80 eighth graders in Experiment 1, 56 eighth graders in Experiment 2, and 42 ninth graders in Experiment 3. Results of two of the experiments supported the first major hypothesis, thus showing that internalization of schematic filmic codes is possible and leads to improved performance on related transfer tasks and ability tests. Aptitude-treatment interactions emerged in all three experiments, as expected by the second major hypothesis. It was concluded that filmic modeling of schematic operations can lead to their internalization, thereby improving the ability of low scorers to use the operations as covert mental skills. Verbal ability was not found to be a necessary mediator in this kind of learning.

Bruner postulated on a number of occasions (Bruner, 1961; Bruner et al., 1966) that for communication systems, tools, or media to be effective, they must produce appropriate internal counterparts in their users' minds. Berlyne (1965) stated similarly that signs appear to have a dual function: they serve both for overt *communicational* and for covert *representational* purposes. More recently, Olson (1972) suggested that while different forms and media of instruction convey knowledge that maps onto a common knowledge system, they diverge as to the mental skills they cultivate. These ideas are far from being new in the realm of language. In fact, much of the verbal training given to lower-class children is based upon the assumption that improvements in one's communicational capacity are internalized and lead, therefore, to better intellectual functioning (e.g., Blank & Solomon, 1968).

The hypothesis that communication codes can be internalized to be used as "mental tools" need not be limited to language. Media of communication, it is often claimed, utilize different symbolic codes of a nonverbal nature (e.g., Eisner, 1970; Gombrich, 1972). It is difficult to prove that these symbolic codes constitute different grammars in the regular sense. However, it can be shown that they do consist of specific and quite diverse symbols and modes of combining them (e.g., Gordon, 1969). Moreover, some media, particularly film and TV, contain in their symbolic systems specific codes to represent relatively unique transformations in space and time (e.g., slow motion, the zoom of a camera, rotations, etc.).

[1] The research reported here was partly supported by a grant from the American Psychological Foundation and partly by the Israel Institute of Applied Social Research. The author is grateful to Michal Siman-Tov, Deborah Malveh, and Avraham Cohen for their assistance in carrying out the experiments and to Susan Harter at Yale University for her helpful suggestions.

[2] Requests for reprints should be sent to Gavriel Salomon, School of Education, The Hebrew University, Jerusalem, Israel.

Given the recent accumulation of findings pertaining to the role of nonverbal mediation (imagery) in mental development (Piaget & Inhelder, 1971), memory, and problem solving (Arnheim, 1970), one may wonder whether these can be internalized from the communication media to which one is exposed.

This would be one way to examine Bruner's (1961) postulation about the development of internal counterparts to media and to test Olson's (1972) hypothesis about the mentally cultivating function of media and modes of instruction.

The hypothesis we wished to test is that one learns to use covertly in his representational system certain operational schemes which he encounters as part of a medium's language of communication. Thus, for example, we would expect someone to better visualize a "slowed" operation after intensive exposure to films which show "slowed down" movements. Once such a scheme is internalized, it should serve as a mediating mechanism which facilitates performance in relevant problem situations (Lesser, 1972). The "internal counterpart" to which Bruner and Berlyne refer would then become exactly that: What is originally a part of a medium's language is now transformed into covert visualization for future use.

What is the nature of this process? The analogy with language has its limits. Although some imitation may take place in early childhood, the acquisition of language is characterized by interaction with others (e.g., Brown & Bellugi, 1964) and by the fact that the child both *encodes* and *decodes* verbal messages. This does not take place in other media where the child serves mainly as a decoder.

Nevertheless, there are several indications that even under such conditions, internalization of communication codes is possible. Most clearly, there is the possibility of *imitation.* Although Bandura (e.g., 1965) limited himself to studying the imitation of human models (although some were put on film), there is no reason why a learner could not imitate nonhuman objects and their "behaviors." Piaget (1962) is quite explicit about this possibility and provides empirical observations to support it. The imitation and internalization of operations are, according to him, a major element in the generation of schematic images and are part of one's developing intelligence. Thus, the fact that one does not use the scheme as an encoder or interact with others by means of such a code may not prevent its internalization.

However, modeling overtly a particular transformation, which has the potential of becoming an internalized schematic mediator, may have differential effects depending on the learners' initial mastery of functionally equivalent mediators. Instances in which learners were given ready-made mediators resulted often in interference (Bruner, 1961; Gentile, Kessler, & Gentile, 1969; Jensen, 1967). However, interference implies that the needed mediators are already available in the learner's repertoire. Improvement, rather than interference, can be expected when the learner does not initially master the explicitly modeled skill, that is, when such a presentation is sufficiently novel and carries some promise for better performance. On the other hand, already skillful learners would be expected to improve from a presentation which is less explicit and which calls upon a mediator they have already mastered to some degree. It can thus be hypothesized that initial ability interacts with the explicitness with which a particular transformation is presented.

Referring to the issue of communication codes which represent schematic transformations in space or time, we reason as follows. If a medium such as film does present explicitly a particular transformation which *could* become a *covert* representational scheme as a consequence of imitation and internalization, then initially less skillful learners would be expected to demonstrate better mastery of that representational transformation. More skillful learners would be expected to experience interference and to demonstrate depressed performance. The latter could, however, benefit more when given the opportunity to provide that representational transformation on their own.

In the three experiments reported below, the above hypotheses were tested. Explicit-

ness of presenting the transformation to be internalized was operationally defined as three points along a continuum:

1. Maximal explicitness. Films are shown that present a transformation, such as singling out an item through gradually "zooming in" on it (Experiments 1 and 2) or "laying out" a solid object (Experiment 3), in fullest detail. That is, there is an initial state, a transformation it undergoes, and its resultant state (S_1–tr–S_2). In this condition, the symbolic code of the medium is used to its fullest; learners are expected to imitate and to internalize the presented transformation. We labeled it the *modeling* condition.

2. Partial explicitness. Slides, instead of film, are shown which have only the initial and final states, leaving out the transformation leading from the former to the latter (S_1–S_2). In this condition, learners are expected to covertly supply the linking transformation. Consequently, this was labeled the *short-circuiting* condition.

3. Minimal explicitness. A slide is shown which depicts only the initial state (S_1) and which leaves out the transformation and the resultant state (S_2). Given appropriate instructions, learners are expected to covertly activate and apply the appropriate transformation in order to come up with the final situation. This was called the *activation* condition.

The two filmic transformations selected for the experiments—singling out items through "zooming in" on them and "laying out" of objects—were chosen because they met two necessary criteria: (*a*) they were part of the medium's range of schematic codes and (*b*) there was reason to believe that these transformations can be internalized and thus enrich one's covert mediational system.

Experiment 1

The filmic code of "zooming in" on details that are embedded in a wider context was selected as the transformation to be modeled by film. This coded transformation was selected because we had good reason to believe that "zooming in" on details is analogous to the analytic process of singling out discrete components of a rich stimulus. Thus, there was reason to hypothesize that "zooming in" has the potential of serving as a model, which could be imitated and internalized. It should, as a consequence, lead to an improved ability of cue attendance, as studied by Sieber and Lanzetta (1966) and Salomon and Sieber-Suppes (1972).

Given the hypothesis presented earlier, the experiment called for four experimental conditions: modeling (maximal explicitness), short-circuiting (partial explicitness), activation (minimal explicitness), and a no-treatment control group. Since, however, aptitude-treatment interactions were expected, rather than main effects, subjects' initial cue-attendance ability had to be pretested.

Method

Stimuli. The *modeling* stimulus condition consisted of three, super 8-millimeter films. Each of these depicted one of Breughel's paintings ("Children in the Playground," "Proverbs," and "Winter in the Village"). The camera "zoomed in" on details in a random sequence (e.g., a particular child playing, a woman in a window, etc.), and "zoomed out" again. This was repeated 80 times in each film. The *short-circuiting* stimulus condition consisted of three series of slides, corresponding to the three films in the modeling condition. In each series, a slide depicting one of the three paintings was shown first, followed by another slide depicting a detail, then the whole painting again, followed by another detail. This was repeated for 80 times in each series. The details that were singled out were identical to the ones shown in the modeling films, with the same random order of presentation and length of exposure. The *activation* condition consisted of only three slides, each showing one of the three paintings.

Procedure. Subjects in all three treatment conditions were required to report in writing 80 details they had noticed in each film or slide. This was repeated three times (once for each painting) so that each subject reached the criterion of reporting 240 details altogether.

Subjects in each treatment condition, with the exclusion of the control group subjects, were seated together in one room, given an introduction and examples, and then worked individually; that is, each subject reported in writing, the 80 details he noticed upon viewing a film or slide. Once finished, an experimenter read over the subject's list of noticed details. In case a detail was mentioned more than once, the subject was requested to replace it. Time to criterion varied: subjects in the modeling and the short-circuiting groups worked according to the speed determined by the presentations; the subjects in the activation group worked as long as needed, and the slide was projected until the last subject finished his task.

Subjects. 80 eighth graders, all from one school, participated in the study. They were randomly assigned to four groups ($n = 20$): modeling, short-circuiting, activation, and a pretest- and posttest-only control group. There was an equal number of males and females in all groups.

Measures. A pre- and posttest measure of cue attendance was group administered to all subjects. The purpose of the cue-attendance *pretest* was to serve as an aptitude measure upon which posttest scores could be regressed, given the hypothesis that treatments and aptitude interact with each other.

The cue-attendance[3] pre- and posttest were similar, though not identical in nature. Subjects, seated in a group, were shown a slide depicting a rather complex visual montage of items and were asked to report in writing, the maximum number of items they could notice. There was a time limit of seven minutes.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF PRE- AND POSTTESTS FOR ALL GROUPS

Group	Cue attendance			
	Pretest		Posttest	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*
Modeling	13.12	6.90	31.3[c]	8.60
Short-circuiting	11.45	6.07	25.1[b]	7.40
Activation	13.05	4.17	32.8[c]	9.95
Control	12.70	6.65	16.7[a]	6.50

Note. Means which are significantly different from each other ($p < .05$) have different superscripts.

Results

As shown in Table 1, all four groups had quite similar scores on the pretest. However, significant group differences were found ($F = 4.52$, $df = 3/76$, $p < .01$, one-way analysis of variance) on the posttest. All three experimental groups reported noticing significantly more items on the cue-attendance posttest than the control group, indicating an accord with previous studies (Salomon & Sieber-Suppes, 1970; Sieber & Lanzetta, 1966) that this skill is modifiable.

In addition, the posttest performance of the modeling group was not significantly better than that of the activation group, but both were better than the short-circuiting group. This seems to suggest that on the average, much explicitness or very little explicitness is more effective than partial explicitness.

TABLE 2

LINEAR REGRESSION COEFFICIENTS OF POSTTEST SCORES PREDICTED FROM PRETEST SCORES FOR ALL GROUPS

Group	*r*	*b* coefficient	*t*[a]
Modeling	−.46	−.572	2.21*
Short-circuiting	+.28	.341	1.22
Activation	+.52	1.24	2.78**
Control	+.48	−.469	2.50*

[a] All *t* values are the difference from zero.
* $p < .05$.
** $p < .01$.

More directly related to our hypothesis concerning an aptitude-treatment interaction is the examination of regression lines within each treatment condition. Thus, cue-attendance posttest scores were regressed on pretest (aptitude scores), as presented in Table 2.

The analyses revealed a significant interaction between the initial cue-attendance aptitude and the conditions of modeling and activation ($t = 3.15$, $df = 36$, $p < .001$, see Figure 1).

In agreement with the hypothesis, low scorers on the aptitude test appeared to have benefited relatively little from the activation

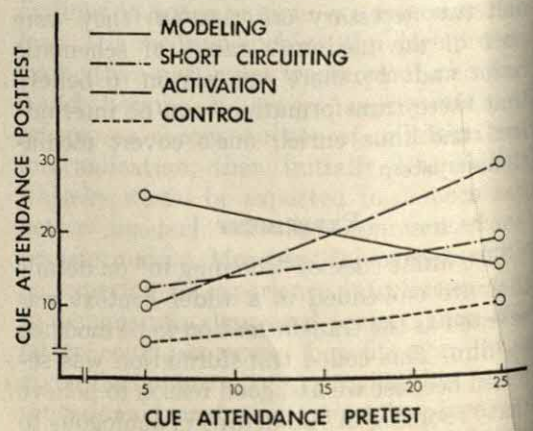


FIGURE 1. Cue-attendance posttest scores regressed on pretest scores for each group.

[3] More information pertaining to this measure can be found in Salomon and Sieber-Suppes (1972).

condition. This condition, as hypothesized, called upon mediators that low-aptitude scorers were not expected to master initially. However, they have profited far more from the modeling condition which, as we have reasoned, provided a medium-generated code that could be imitated and internalized. While low-aptitude scorers appeared to learn from the highly explicit filmic presentation of the transformation, initially high scorers performed *less* well following modeling and far better following the activation condition. It appears that the very explicit modeling films offered a mediating transformation which interfered with the already developed mediational capability of those subjects. They did profit, however, from a condition which called upon mediators that they had mastered already.

The short-circuiting condition appeared to benefit low- and high-aptitude scorers alike.

Discussion

The results of this experiment lend tentative support to the two major hypotheses:

1. A filmic transformational scheme, in this case "zooming in" and "zooming out," is learnable; it can be used *covertly* in a similar task with new material (the cue-attendance posttest).

2. There is a negative relationship between the explicitness of presenting the transformational scheme (i.e., the visual presentation of the operation) and the learner's initial ability to execute it covertly on his own. The less able he is, the more he profits from such modeling.

Although results were as expected, there is nothing in the data to suggest how the learning took place when one "internalized" the filmic scheme. If it was imitation, then it should not have been limited to only one scheme, given that our films also used other operational schemes at the same time. For instance, the *order* in which details were singled out was such an additional scheme, and subjects would be expected to imitate *it* as well. Second, if it was not just simple visual imitation, and the learning was mediated by internal verbalization, then induced verbalization should lead to even more improvements. Experiment 2 was designed to partially replicate experiment 1 and to shed light on the above questions.

Experiment 2

This experiment partially replicated the former one. However, one experimental variable was added, namely, induced overt labeling of the items that were singled out by the camera (in the modeling condition) or by the subject himself (in the activation condition). It was reasoned, following Kendler and Kendler (1968) that if verbalization plays a mediating role in learning from the modeling films, then induced labeling should improve posttraining performance. Low-verbal-ability subjects were expected to gain more from such labeling than more verbally able subjects who apparently were more likely to use verbal mediation on their own. However, if this kind of learning does *not* rely upon verbal mediation, then, it was reasoned, induced verbalization would have no effect on learning.

In the present experiment, the short-circuiting condition was dropped, leaving only the modeling and activation conditions. These are the two extreme ends of the continuum of modeling a covert skill through the use of filmic operational schemes. This resulted in a 2 × 2 (Modeling versus Activation × Verbalization versus No Verbalization) factorial design.

Method

Stimuli. There were two stimulus conditions which were identical to the modeling and activation conditions of the previous experiments. Half of the subjects saw the films which modeled the operation of singling out details from a rich display, using the "zooming in" technique (the modeling condition). The other half saw three slides of these paintings on which the films were based (the activation condition).

Procedure. There were four treatment groups.

1. Modeling–verbalization group. Subjects saw the three modeling films. While viewing them, one subject was randomly called upon on each trial to label aloud the detail on which the camera was then "zooming in." Subjects saw the films in groups of seven (half the size of the group). It was *assumed* that all subjects were busy labeling the details to themselves, since no one could know whose turn it would be next to label aloud. The criterion to be reached by each subgroup of seven subjects

was 42 labeled details per film (6 reported details per subject). Each child was reinforced for his labeling. The whole training period lasted for about three hours.

2. Modeling–nonverbalization group. Subjects were run in the same fashion as above, with the exclusion of verbalization. No verbal responses were required, and no reinforcements were given. Subjects were told to "notice what exactly the camera does and to note to themselves the details they observe." The films and the criterion were the same (criterion was actually set by the number of "zoom ins" per film).

3. Activation–verbalization group. The same procedure as in the first group was maintained. However, instead of being exposed to the films which singled out details for the subjects, there was a static slide and subjects had to report noticing the same number of details from the slide, as in the first group.

4. Activation–nonverbalization group. This group served actually as a control group since it received the same instructions as the other groups, but the subjects watched the static slides in silence. Instructions were those given to the second group.

Subjects. 56 eighth graders were randomly chosen from two classrooms. These were then randomly assigned to the four groups ($n = 14$). Within each group, the 14 subjects were again randomly divided into two subgroups of 7 each. This was the unit with which the experimenters worked.

Measures. The number of aptitude and posttraining measures was increased from Experiment 1. The reasons for increasing the number of pretest aptitude measures were, first, that cue attendance, as measured by us, was a rather limited aptitude with unknown relations to other aptitudes. Hence, the Embedded Figures Test (measuring Field Dependency, Witkin, 1964) was added. Second, since we hypothesized verbal ability to interact with treatments, an appropriate test had to be added. Posttest measures were increased to enable us the examination of what else might be internalized from the films.

There were three pretest-aptitude measures administered in a random order one day prior to experimentation: (*a*) *cue-attendance pretest*, very similar in nature and requirement to the tests used in Experiment 1; (*b*) Witkin's Embedded Figures Test (to measure Field Dependency); and (*c*) an Israeli Standardized Verbal Ability Test (MILTA), for which norms based on national samples are available.

The posttest measures, administered in a random order one day following the end of experimentation, were as follows: (*a*) *A cue-attendance posttest*, which was based like the previous cue-attendance tests, on a complex and rich visual display, from which subjects had to single out details and report them in writing. (*b*) *An organization measure*, extracted from the cue-attendance test, which indicated the extent to which details were noticed in some spatial order. This measure was included since we hypothesized that subjects might imitate not only the act of "zooming in" but the order (in fact, random order) in which items were "zoomed in" on. Whenever a subject reported noticing a detail which was spatially adjacent to a formerly reported detail, he received one point. Subjects were not told to report details in any order, nor was the question of order mentioned.

Results

Means and standard deviations of the four groups on the three pretests and the two posttest measures are presented in Table 3.

It should be noted that cue-attendance pretest scores are higher than those in Experiment 1 due to richer and more readily available details in the stimuli in Experiment 2. Posttest mean scores are lower in the postest than in the pretest because of more difficult stimuli.

Factorial 2 × 2 analyses of variance failed to produce statistically significant F ratios for either of the two main effects or

TABLE 3

MEANS AND STANDARD DEVIATIONS FOR EACH GROUP

Group	MILTA		Embedded Figures Test		Cue attendance				Organization	
					Pretest		Posttest			
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
Modeling										
With verbalization	92.8	13.9	6.5	4.7	50.9	10.9	35.9	2.7	36.2	15.4
No verbalization	92.7	12.8	6.5	4.5	48.6	9.6	34.0	6.2	37.5	18.7
Activation										
With verbalization	87.4	10.7	6.8	4.0	48.8	9.5	33.7	6.3	35.3	17.8
No verbalization	89.1	11.9	6.5	3.2	50.9	10.8	32.6	6.5	33.7	10.2

Note. Abbreviation: MILTA is an Israeli Standardized Verbal Ability Test.

for the interactions between them. The failure to detect significant differences between the treatments and the control group disagrees with the results of Experiment 1, although the ages of the subjects in the two experiments were very similar, and the stimuli were identical. The only major difference between the two experiments was in the level of training criterion that each subject had to reach individually. In Experiment 1, each subject had to report in writing 80 details that were noticed per film (or slide) and 240 details altogether. In Experiment 2, the individual criterion was only 6 verbally reported details per film (or slide), and 18 altogether, per subject. Thus, it would be reasonable to postulate that in Experiment 2 the criterion to be reached was apparently far too low to produce any observable difference which reaches the needed level of significance.

Our main concern was, however, with aptitude-treatment interactions (ATI). To detect such interactions, posttest scores were regressed on aptitude scores for each group. The intercorrelations among the measures are given in Table 4.

What concerns the differential relationship between pretest and posttest cue-attendance is that we find an interaction between the modeling–no-verbalization and the activation–no-verbalization (control) conditions (difference between regression slopes: $t = 3.30$, $df = 26$, $p < .01$). Such an interaction does *not* take place, however, between modeling and activation when overt verbalization is required.

In other words, whenever subjects are required to activate covertly the necessary process of singling out details *or* when they are to overtly label the singled-out details, initially good cue attenders benefit most. When, on the other hand, no labeling is required and the process of singling out details is explicitly modeled, poor cue attenders benefit most (Figure 2).

TABLE 4

INTERCORRELATIONS BETWEEN MEASURES SEPARATELY FOR EACH GROUP

Measure/group	MILTA	Field Dependency	Cue attendance		Organization
			Pretest	Posttest	
MILTA					
Modeling – verbalization	—	.875**	.560*	.271	.605*
Modeling – no verbalization	—	.528*	.552*	−.180	.800*
Activation – verbalization	—	.531*	.523*	.730*	.153
Activation – no verbalization	—	.653*	.630*	.580*	−.063
Field Dependency					
Modeling – verbalization		—	.271	.334	.686*
Modeling – no verbalization		—	.355	−.088	.516*
Activation – verbalization		—	.266	.384	.073
Activation – no verbalization		—	.377	.377	−.273
Cue-attendance pretest					
Modeling – verbalization			—	.403	.356
Modeling – no verbalization			—	.007	.372
Activation – verbalization			—	.411	.145
Activation – no verbalization			—	.549*	.311
Cue-attendance posttest					
Modeling – verbalization				—	−.108
Modeling – no verbalization				—	−.082
Activation – verbalization				—	.453
Activation – no verbalization				—	.566*

Note. $n = 14$ in each group. Abbreviation: MILTA is an Israeli Standardized Verbal Ability Test.

* $p < .05$.

** $p < .01$.

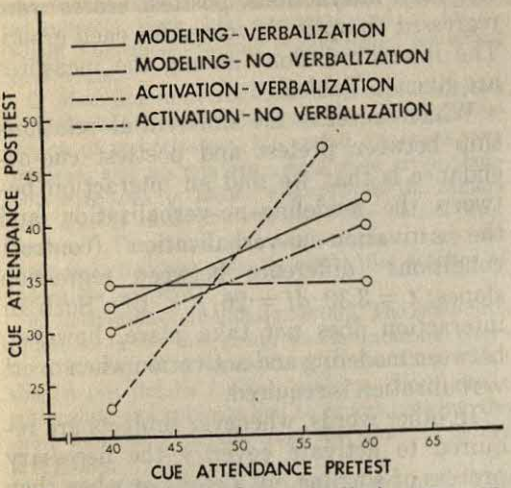


FIGURE 2. Cue-attendance posttest scores regressed on pretest scores for each group.

Another interaction is observed when cue-attendance posttest scores are regressed on MILTA pretest scores. Here, the interaction is also between the modeling–no-verbalization and the activation–no-verbalization conditions. While initially, less verbally able subjects benefit most from the modeling condition, they benefit far less from the activation condition. The converse is true for the more verbally able subjects (difference between regression slopes: $t = 2.65$, $df = 26$, $p < .05$; Figure 3). It should be noted, however, that the correlation between MILTA and cue-attendance pretest scores is .474 ($N = 56$, $p < .01$); hence, the similarity of the two aptitude-treatment interactions.

It appears that when no labeling is required, initially high-aptitude scorers benefit more when the transformation, which they have mastered already, is called upon (activated) rather than modeled. For the initially low-aptitude scorers, the converse is true. We thus find additional support for our hypothesis concerning the differential effectiveness of explicitly modeling a transformation via a filmic code.

Verbalization, or labeling, usually facilitates the performance of those who are *already* verbally more able, while it does not facilitate the performance of subjects who are less verbally able. This seems to suggest that the latter, who tend also to be poor cue attenders, internalize the transformation and use it covertly as a *nonverbal* mediator. Internal verbalization seems therefore to be an unnecessary mediator for such performance to take place.

Other interactions emerge when organization scores are regressed on MILTA or Field-Dependency scores (overall correlation between the two is .692), as shown in Figures 4 and 5 ($F = 5.16$ and $F = 5.62$, respectively, $p < .05$). The interesting point to note is that low MILTA scorers and high Field Dependency ones scanned the visual field in a much less organized way, following the modeling treatments, than similarly low scorers following activation conditions. On the other hand, low scorers on the two aptitude tests, who were *not* exposed to the model (which as it will be recalled, displayed a random order of "zooming in" on details) showed better organization in their scanning, which did not differ significantly from that of those with high aptitude. It should also be noted that *high*-aptitude subjects scanned the field in a much more organized fashion following the modeling conditions than those high-aptitude subjects in the activation conditions.

Whether this shows some kind of overcompensation on the side of the latter subjects in face of the unorganized model remains an open question. It is the case, how-

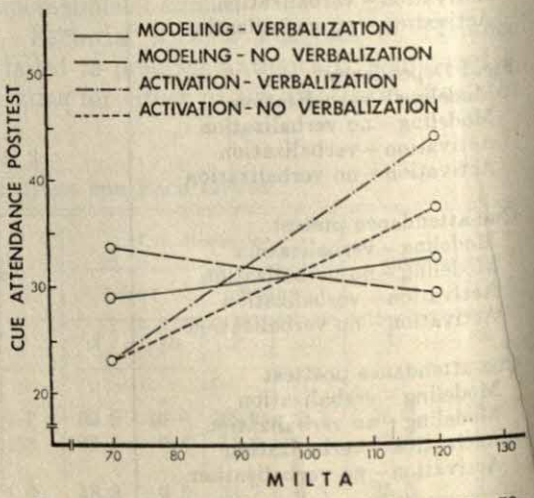


FIGURE 3. Cue-attendance posttest scores regressed on MILTA pretest scores for each group. (Abbreviation: MILTA is an Israeli Standardized Verbal Ability Test.)

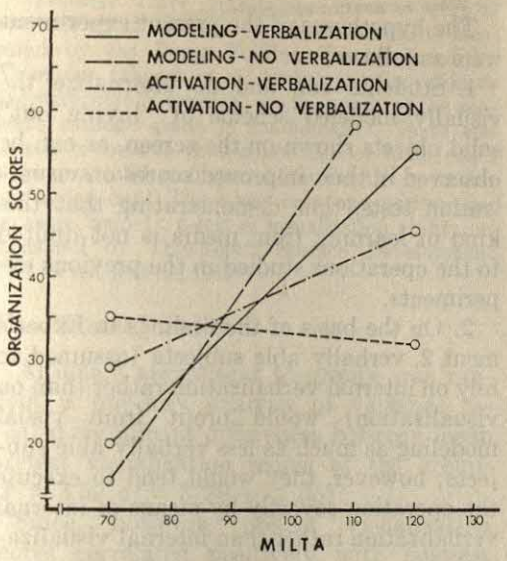


FIGURE 4. Organization scores regressed on MILTA pretest scores for each group. (Abbreviation: MILTA is an Israeli Standardized Verbal Ability Test.)

ever, that high MILTA scorers, following the modeling conditions, noticed significantly fewer details (as measured by the posttest) than high MILTA scorers who were exposed to the activation conditions. Comparing the cue-attendance performance of the 10 highest scorers on MILTA within the modeling conditions with the 10 highest MILTA scorers in the activation conditions shows that the latter perform a bit better than the former ($t = 1.8, df = 18, p < .10$). The *opposite* occurs when we compare the *organization* scores of these two subgroups ($t = 2.01, df = 18, p < .05$). Thus, it appears that when exposed to a model which (*a*) shows an operation with which these subjects are familiar and (*b*) displays another operation (random order of scanning) that apparently disagrees with their more orderly style, these subjects put more energy into imposing structure on their responses than into producing a large number of such. Therefore, their organization scores correlate highly with verbal ability when in the modeling conditions (.70 with MILTA and .72 with Field Dependency) and not at all with their posttraining cue-attendance performance (−.08). The exact opposite pattern takes place in the activation conditions (see Table 4).

Discussion

In general, then, apparently due to training criteria that were too low, no clear effects of either modeling the schematic operation or of induced verbalization were found. The assumption that *all* subjects in the verbalization groups engage in spontaneous labeling while expecting to be called upon, appears to be questionable. It is more likely that the low-verbal-ability subjects engage in labeling only *after* being called upon. Therefore, the criterion of 18 labels which they had to provide was far too low for them. On the other hand, the aptitude-treatment interactions show that learning—in terms of internalization of the modeled operation—takes place. However, this is restricted to the low-aptitude scorers, while the high scorers either experience interference or try to rely on their high verbal ability to overcome the disagreement between the operations on which they typically rely and the ones explicitly modeled for them. What the interactions with MILTA as predictor suggest is that *some* subjects (particularly *low*-verbal-ability scorers) imitate the visual schematic operation quite directly, something that is seen in the number of details they report and in the *reduced* organization

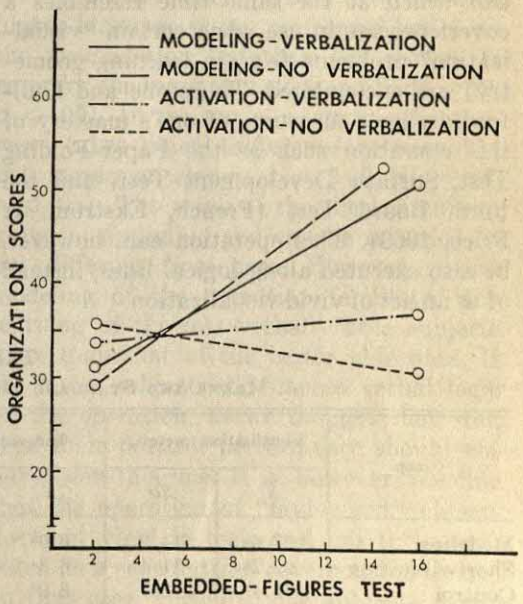


FIGURE 5. Organization scores regressed on Embedded Figures Test pretest scores for each group.

of their responses. Other subjects (particularly the more verbally able ones) do not just imitate the visual schematic operation but internalize it through verbal mediation. This in turn causes them to produce fewer responses but to chunk and organize them better.

Experiment 3 was partly designed to shed more light on this question. It was hypothesized that if more verbally capable subjects do impose logic on the operation which they internalize, then their learning should not be debilitated. This, however, requires one necessary condition, namely, that the schematic operation to be internalized *can* be executed by means of logical operations and not only by means of dynamic images, as is apparently the case with the "zoom in" operation.

Experiment 3

In this experiment, we returned to the design of Experiment 1, leaving out only the activation condition and choosing to model or short-circuit an operation which is far more novel to subjects than the one modeled before. The schematic operation modeled here was that of "laying out" solid objects. This again is an operation which is within the filmic range of schematized operations and which at the same time resembles a covert process in use when certain "visualization" problems (e.g., in learning geometry) are encountered. Thurstone and Guilford designed measures of one's mastery of this operation such as the Paper-Folding Test, Surface Development Test, and the Form Board Test (French, Ekstrom, & Price, 1963). This operation can, however, be also executed along logical lines, instead of as an act of vivid visualization.

The hypotheses of the present experiment were as follows:

1. Students can and do internalize the visually modeled scheme of "laying out" solid objects shown on the screen, as can be observed in their improved scores on visualization tests, thus demonstrating that this kind of learning from media is not limited to the operations studied in the previous experiments.

2. On the basis of the findings in Experiment 2, verbally able subjects (assumed to rely on internal verbalization rather than on visualization) would profit from visual modeling as much as less verbally able subjects; however, they would tend to execute the operation covertly by means of internal verbalization rather than internal visualization.

Method

Stimuli. There were two treatment stimuli: modeling and short-circuiting. The modeling stimulus consisted of a 15-minute film in which five solid objects appeared and then gradually were laid out, side after side, to produce a two-dimensional plan of the object. Once an object was laid out, it gradually folded up again. The short-circuiting stimuli consisted of a series of 10 slides (5 pairs). The first slide in a pair showed the solid object while the subsequent member of the pair displayed the same object in a laid out position.

Procedures. There were three experimental groups: modeling, short-circuiting, and a control group. The modeling group was given a general introduction and then shown the film once. It saw the same film again on the next day and a third time on the third day. No responses were required. The short-circuiting group received the same treatment but with the slides instead of the film. The control group served as a pretest- and posttest-only group. No treatment was given.

Subjects. The subjects were 42 ninth-grade students in a vocational school. All were males. They were randomly assigned to the three groups (n = 14).

TABLE 5

Means and Standard Deviations for Each Group

Group	Visualization (pretest)		Language studies		Mathematics		Visualization (posttest)	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
Modeling	62.6[a]	14.9	6.0[a]	.59	6.0[a]	0.8	68.0[c]	9.5
Short-circuiting	62.2[a]	11.4	6.3[a]	.68	6.1[a]	1.4	63.5[b]	16.7
Control	71.4[b]	13.7	6.0[a]	.79	5.6[a]	1.0	57.0[a]	25.5

Note. Means which differ significantly from each other ($p < .05$) have different superscripts.

Measures. Three pretests were given to serve as aptitude measures: (*a*) visualization ability, measured by the Paper Folding Test designed by Thurstone (French et al., 1963); (*b*) the subject's average grade in language studies; and (*c*) the subject's average grade in mathematics. There was reason to believe that either one of them could serve as a valid predictor of learning the operation of "laying out." There was only one posttraining measure, namely, a test of visualization ability, as measured by Thurstone's Surface Development Test (French et al., 1963).

Results and Discussion

Although assignment to groups was random, a one-way analysis of variance revealed significant differences between mean pretest visualization scores of the groups ($F = 3.52, df = 3/39, p < .05$).

Since, however, pretest visualization scores correlated positively with posttest visualization scores in all groups and since linear requirements were met, analysis of covariance was used. The Scheffé post hoc comparison showed that the mean posttest visualization scores of the modeling group were significantly higher than that of the short-circuiting group and that both means were significantly higher than that of the control group (Table 5). Thus, the results appear to be in agreement with those of Experiment 1.

TABLE 6

INTERCORRELATIONS BETWEEN MEASURES FOR EACH GROUP

Measure/group	Visualization pretest[a]	Language grades	Mathematics grades	Visualization posttest[b]
Visualization pretest				
Modeling	—	.54*	.41	.29
Short-circuiting	—	.20	.24	.35
Control	—	.30	.13	.31
Language grades				
Modeling	—	—	.12	−.36
Short-circuiting	—	—	.44	−.19
Control	—	—	.14	.63*
Mathematics grades				
Modeling	—	—	—	.38
Short-circuiting	—	—	—	.21
Control	—	—	—	.28

Note. $n = 14$ in each group.

[a] The visualization pretest consisted of the Paper-Folding Test.

[b] The visualization posttest consisted of the Surface Development Test.

* $p < .05$.

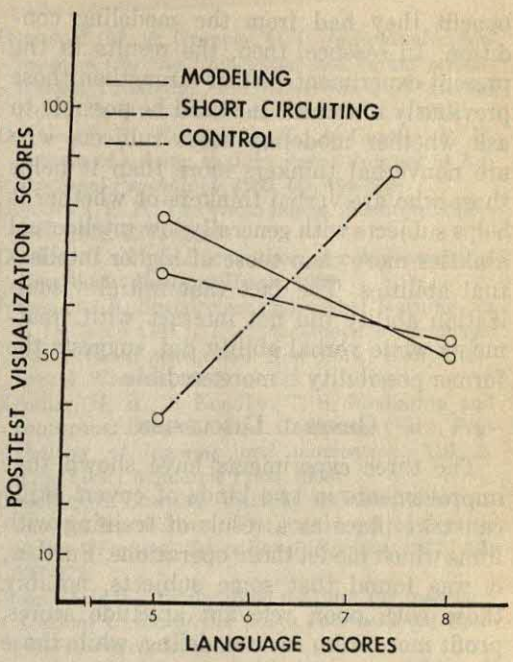


FIGURE 6. Visualization posttest scores regressed on language grade for each group.

When correlational analyses were done (Table 6), an interaction emerged between the subjects' grades in the language studies and their posttest visualization scores (Figure 6). In the control group, subjects with higher language scores performed better on the visualization test than low scorers (t ratio for the regression slope is 2.8, $df = 12$, $p < .05$). In the modeling group, the converse takes place (difference between the two regression slopes: $t = 2.08$, $df = 24$, $p < .05$). The slope of the short-circuiting group is negligible and can be regarded as not differing from zero. However, visual modeling of the operation facilitated the learning of the *less* verbally able subjects more than that of the better able ones. If the latter subjects did impose verbal logic on the operation, as we thought they did, then their posttest performance should not have been this low. It is, however, possible that the operation of "laying out" objects does not yield to logic and has to be executed as a schematized covert visual image. In that case, verbally able subjects would try, in vain, to utilize logic; hence, the little

benefit they had from the modeling condition. In essence, then, the results in the present experiment further strengthen those previously obtained. It would be possible to ask whether modeling helps subjects who are nonverbal thinkers more than it helps those who are verbal thinkers or whether it helps subjects with generally low intellectual abilities more than those of higher intellectual abilities. The fact that initial visualization ability did not interact with treatments while verbal ability did, suggests the former possibility is more credible.

General Discussion

The three experiments have shown that improvements in two kinds of covert skills can take place as a result of training with films which model those operations. Further, it was found that some subjects, notably those with poor, relevant aptitude scores, profit more from such modeling, while those with high scores are hindered in their performance. The latter subjects profit more when asked to execute the operation covertly on their own. It was also observed that vicarious learning, in the sense that a subject executes covertly a response while another verbalizes it (Experiment 2), is not the same as when he himself has to act out the response (Experiment 1). Hence, when training criterion is lowered, learning from a visual model is reduced. Finally, the hypothesis that some subjects are better off internalizing the model through nonverbal mediation, received indirect support.

However, one may wonder whether all that has been shown here is perhaps no more than another case of learning from visual displays. This, obviously, is not very novel in light of what both daily experience and research have repeatedly shown. Gagné and Gropper (1965) and Travers (1970), to mention only a few, have repeatedly shown that learning from visual displays is not only possible but even relatively more powerful than verbal instruction.

There is, however, a major and very essential difference between the usual kind of studies dealing with visually based instruction and the ones reported here. Clearly, the operations displayed on the screen by us cannot be termed signs or symbols since they do not serve as completely arbitrary representations of operations but are the operations themselves. In this respect, they are *iconic* representations in the sense that we assume that they have something in common with the covert operations which they "externalize." That is, one can assume that they resemble the operations which the learners should have executed or actually did execute covertly. The operations utilized here are not analogous to grammatical forms used in language or to verbal concepts. Nevertheless, they differ markedly from the usual information displayed in films or slides. It was not the details of Breughel's paintings nor the structure of the solid objects which was to be learned from these slides but *rather the schematic operations of singling out details or "laying out" solid objects*. Moreover, the question we posed was not whether visual information is better for instructional purpose than verbal information or modeling is better than "autonomous" learning; but rather, whether such schematic operations as sampled here, which are part of a medium's unique range of communication codes, are internalizable *as schematic* operations and, hence, whether they are used as *covert schemes*.

One may ask whether exposure to media, such as print, film, TV, and the like leads to the cultivation of covert representational systems, as Olson (1972) or Bruner et al. (1966) claim. This, we hypothesized, is a matter of internalizing both the codes, conventions, and schematic ways of representing something, which are unique to media, and their covert use as part of one's representational system. The experiments reported here were a first attempt at studying whether such internalization is at all possible, and if so, by whom.

The internalization of operations and their use in a representational capacity are an important issue in Piaget's theory. However, he strongly emphasizes the importance of manipulatory learning rather than observational learning. Nevertheless, while discussing imitation (Piaget, 1962), observations of operations appear to play as important a role in his theorizing as manipu-

lation. But this may not be an all-or-none question. *Logical* operations, of the kind studied by Piaget, may be learned solely by manipulation (Wohlwill, 1970). Not so with other operations which, as the ones studied by us, do not follow necessarily any agreed-upon logic and are rather conventional schemes.

Educationally speaking, what our studies hint at is that certain mental skills may be adopted from communication media and thus be used to expand one's range of covert skills. The question, then, is not whether this is a "better" mode of instruction, but whether one can use visual media not just to acquire "knowledge *that*" but also "knowledge *how to*," particularly for those learners who appear to have difficulties with other and more common types of instruction.

REFERENCES

Arnheim, R. *Visual thinking*. Berkeley: University of California Press, 1970.

Bandura, A. Behavioral modification through modeling procedures. In L. Krosner & C. P. Ullman (Eds.), *Research in behavior modification*. New York: Holt, Rinehart & Winston, 1965.

Berlyne, D. E. *Structure and direction in thinking*. New York: Wiley, 1965.

Blank, M., & Solomon, F. A tutorial language program to develop abstract thinking in socially disadvantaged preschool children. *Child Development*, 1968, **39**, 379–389.

Brown, R., & Bellugi, U. Three processes in the child's acquisition of syntax. *Harvard Educational Review*, 1964, **34**, 133–151.

Bruner, J. S. The act of discovery. *Harvard Educational Review*, 1961, **31**, 21–32.

Bruner, J. S., et al. Studies in cognitive growth. New York: Wiley, 1966.

Eisner, E. W. Media, expression, and the arts. In G. Salomon & R. E. Snow (Eds.), *Commentaries on Research in Instructional Media, Viewpoints*, 1970, **46**, 5.

French, J. W., Ekstrom, R. B., & Price, L. A. *Manual for Kit of Reference Tests for Cognitive Factors*. Princeton, N.J.: Educational Testing Service, 1963.

Gagné, R. M., & Gropper, G. L. *Individual differences in learning from visual and verbal presentations*. Pittsburgh, Pa.: American Institute for Research, Studies in Filmed Instruction, 1965.

Gentile, J. R., Kessler, D. K., & Gentile, P. K. Process of solving analogy items. *Journal of Educational Psychology*, 1969, **60**, 494–503.

Gombrich, E. H. The visual image. *Scientific American*, 1972, **227**, 82–97.

Gordon, G. N. *The languages of communication*. New York: Hastings House, 1969.

Jensen, A. R. Social class and verbal learning. In J. DeCecco (Ed.), *The psychology of language, thought, and instruction*. New York: Holt, Rinehart & Winston, 1967.

Kendler, H. H., & Kendler, T. S. Mediation and conceptual behavior. In J. T. Spence (Ed.), *Psychology of learning and motivation*. Vol. 2. New York: Academic Press, 1968.

Lesser, G. S. Learning, teaching and television production for children: The experience of Sesame Street. *Harvard Educational Review*, 1972, **42**, 232–272.

Olson, D. Is the instructional medium the message? *Abstract Guide of the XXth International Congress of Psychology*, 1972.

Piaget, J. *Play, dreams and imitation in childhood*. New York: Norton, 1962.

Piaget, J., & Inhelder, B. *Mental imagery in the child*. London: Routledge & Kegan Paul, 1971.

Salomon, G. Cultural differences in reading and understanding geographic maps. Paper presented at Annual Meeting of the American Educational Research Association, Chicago, February 1968.

Salomon, G., & Sieber-Suppes, J. E. Learning to generate subjective uncertainty: The effects of training, verbal ability and stimulus structure. *Journal of Personality and Social Psychology*, 1972, **23**, 163–174.

Sieber, J. E., & Lanzetta, J. T. Some determinants of individual differences in predecision information-processing behavior. *Journal of Personality and Social Psychology*, 1966, **4**, 561–571.

Travers, R. M. *Man's information system*. Scranton, Penn.: Chandlers, 1970.

Wohlwill, J. F. The place of structural experience in early cognitive development. *Interchange*. 1970, **1**, 13–28.

Witkin, H. A. Origins of cognitive style. In C. Sheerer (Ed.), *Cognition: Theory, research, promise*. New York: Harper & Row, 1964.

(Received July 13, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 512-519

EFFECTS OF NONPARTICIPANT OBSERVATION ON TEACHER AND PUPIL CLASSROOM BEHAVIOR[1]

MICHAEL MERCATORIS AND W. EDWARD CRAIGHEAD[2]
Pennsylvania State University

This study assessed the effects of nonparticipant observation on the behavior of the teacher and mentally deficient pupils in a language development class. Thirty days of classroom behavior were recorded on videotape under an ABAB design. The conditions were: (Condition 1) 10 days, observer absent; (Condition 2) 10 days, observer present; (Condition 3) 5 days, observer absent; and (Condition 4) 5 days, observer present. The subjects were "blind;" they believed they were being recorded only when the observer was in the classroom. The videotapes were observed in random order by "blind" trained observers. Behavioral measures indicated that nonparticipant observation increased the frequency, but not the appropriateness, of teacher–pupil interactions and did not affect the appropriateness of student behaviors.

Recent years have witnessed the extensive application of learning principles in classroom management and teaching methods (see O'Leary & O'Leary, 1972). Data collected by nonparticipant observers in the classroom have been widely used to assess program effectiveness, and until recently the validity of such data has been accepted solely on the basis of high interobserver reliabilities. However, the effects of observer reactivity, observer bias, and methods of computing reliability on the internal and external validity of observational data are currently being questioned empirically (Johnson & Bolstad, 1972; Kazdin, 1973).

The nature of observational reactivity has long been a question in social psychology. Subjective evaluation of the extent of contaminant introduced has ranged from "little effect, if any" (Heynes & Lippitt, 1954, p. 399) to the suggestion that the "patently visible" observer produces changes in behavior that can diminish the generalizability of findings (Webb, Campbell, Schwartz, & Sechrest, 1966). Studies attempting to assess observational artifact and habituation to it have yielded conflicting results. Johnson and Bolstad (1972) noted that both the paradigm of measurement and the nature of dependent variables may contribute to this ambiguity. Further, the nature of reactivity may depend upon the conspicuousness of the observer, individual differences of the subjects, personal attributes of the observer, and the rationale for observation.

Perhaps because of methodological problems, only two studies have directly assessed the effect of observer presence in contrast to observer absence. White (1973), using concealed overhead observation, found that alternating 30-minute periods of observer presence and observer absence had the effect of reducing significantly the activity level of mother–child dyads in the observer-present condition within a laboratory setting. In a second study, using the same equivalent time-samples design with a larger subject pool and a behavioral coding system, he found that observer presence was associated with lower rates of deviant behavior

[1] This study was supported in part by Hospital Improvement Project Grant 51P-70087-3-02 of Social and Rehabilitation Services, U.S. Department of Health, Education and Welfare awarded to Laurelton State School and Hospital, Pennsylvania. Appreciation is expressed to Reeves Power, project director, and Laurelton State School and Hospital for their cooperation in this project.

[2] Requests for reprints should be addressed to W. Edward Craighead, Department of Psychology, Pennsylvania State University, 512 Moore Building, University Park, Pennsylvania 16802.

for older children. Rate of mother deviant behavior and level of instability in rates of social interaction were not associated with conditions of observer presence and observer absence. No comparable observer-present versus observer-absent paradigm has been used in classroom settings.

Two studies have dealt with observer reactivity within educational settings. Gussow (1964) examined the observer-observed relationship in four fourth-grade classrooms in four different schools and concluded that even in situations where the observer is nonparticipant, the observer and the observed were linked in a continuous and developing relationship. Gussow employed retrospective data collection using narrative-descriptive reports that were not germane to the issue of reactivity. In fact, very little data of even a narrative-descriptive nature appeared in his report. Therefore, his conclusions must be cautiously evaluated. Masling and Stern (1969) assessed observer reactivity using trained observers and the teachers and pupils in 23 different classrooms. Two days' observations of each class were broken into 33 five-minute data units which were examined with the hypothesis that if the effects of observers diminished over time, there would be a smaller correlation between the first units of observation and the last than between later units and the last. However, the correlations obtained showed no discernible pattern over time. Two alternative conclusions were proposed: (*a*) Observer influence was negligible; (*b*) the effects of the observers were more complex than foreseen and affected various aspects of teacher and pupil behavior differentially. The authors acknowledged that the failure to include an observer-absent condition precluded the possibility of studying observer effects directly. In addition, by studying habituation the authors may have been unable to detect an effect, since it seems unlikely that a significant habituation effect would have occurred in only two days.

There is no consensus concerning the reactivity of the observational process, particularly within pedagogical settings. Therefore, the present study was designed to answer the question: What effect, if any, does nonparticipant observation have on pupil behavior and on teacher–pupil interactions in the classroom?

Method

Subjects

Subjects were six adult, female residents of a state school for the mentally retarded and their language development teacher, a 23-year-old female, with a bachelor of arts degree in psychology.[3] She was employed as a child care aide and language instructor at the institution and had been conducting language classes for six months at the time of the study.

The students ranged in age from 20 to 44 (M = 31 years), length of institutionalization from 4 to 30 years (M = 15.2 years), and intelligence quotients from 52 to 65 (M = 57.7), the latter falling in the mild to moderately retarded range (Heber, 1961). Only two of the students were considered occasional behavior problems in their residential unit.

Setting

The study was conducted in the students' residential unit, in a room specifically allocated for language classes. The class met from 9:00 to 9:30 a.m., Monday through Friday, and was conducted using the Peabody Language Development Kit, Level I. The teaching procedures were based on behavior modification principles as part of a token economy program at the institution.

A remote-controlled video camera with a wide angle lens and a microphone were placed on the top surface of a metal cabinet in the classroom two weeks before the present study began. The cables from the camera and microphone were led out of the room through a louvered ventilation grate, down a hall, and into an equipment room whose entrance could not be seen from the language classroom. A videotape recorder and monitor were located in this equipment room and were used to record the classroom behavior. The equipment remained in position for the duration of the study.

Procedure

After two weeks were allowed for the class to become adjusted to the presence of the camera and microphone, 30 days of classroom behavior were recorded on videotape using an ABAB design (Baer, Wolf, & Risley, 1968). The conditions were: (Condition 1) 10 days, observer absent; (Condition 2) 10 days, observer present; (Condition 3) 5 days, observer absent; and (Condition 4) 5 days, observer present. The teacher was led to believe that the camera and microphone were operating

[3] Appreciation is expressed to Barbara Salisbury, the teacher, whose understanding permitted the completion of the study.

only when the observer (the experimenter) was in the classroom. Further, she was told that in order for the camera to work the observer must turn it on prior to each class he was recording. This, in fact, was not necessary; the camera was functioning during every class. She was given the set, with a detailed written rationale, that the purpose of the study was to ascertain whether an observer viewing videotapes of a classroom could record behavior as accurately as an observer in the classroom. She was unaware of the true nature of the present study and believed she was participating in an entirely different experiment, the "blind" (reliability) study.

Condition 1 (observer absent, 10 days). Since the observer was not present in the classroom, the teacher was unware of being monitored. She was told that the "blind" study had to be temporarily delayed until the experimenter had obtained and trained a videotape observer.

Immediately preceding or just following the start of each class, the experimenter or his assistant surreptitiously entered the equipment room. Two to five minutes after the beginning of each class, he began to tape the session. He recorded 12 to 15 minutes of classroom behavior, turned off the equipment, and left the room without the teacher's or students' knowledge. There were days when the teacher was aware that either the experimenter or his assistant was in the equipment room; however, she was unaware of the recordings taken during this condition.

Condition 2 (observer present, 10 days). Under Condition 2 the teacher was led to believe that the "blind" study had begun. She was told that an observer would be present in the classroom and would take time-sample recordings of classroom behaviors which would then be compared to ratings made by a videotape observer from tapes of the same class periods. She was informed that 10 days of recordings were necessary for an accurate comparison of the ratings.

Immediately prior to each class, the experimenter entered the classroom and pretended to turn on the camera. He then left the room to get his clipboard, stopwatch, and rating forms and reentered the classroom, taking a seat out of the camera's viewing range. The experimenter maintained the attitude of a nonparticipant observer, avoiding eye contact or interactions with either the teacher or students. Although the experimenter rated classroom behavior for 10 minutes (enhancing the blind given the teacher), the classroom behavior was taped exactly as it was in Condition 1; that is, 12–15 minutes were videotaped by an assistant after the experimenter had been in the classroom for approximately 2 minutes. The experimenter occasionally interacted with the teacher following the class, but he never explained his ratings or made comments or criticisms regarding the class.

Condition 3 (observer absent, 5 days). On completion of Condition 2 the teacher was told that the videotape rater would be making his ratings during the following week and if the reliability between observers was either consistently high or low, then the 10 recorded days would be enough; if not, 10 more days of recorded classroom behavior would be needed. Condition 3 was then conducted in the same manner as Condition 1, while the videotape rater purportedly rated the tapes made under Condition 2.

Condition 4 (observer present, 5 days). On the last day of Condition 3 the teacher was shown percentage agreements, ostensibly classroom and videotape observer reliabilities, that were neither consistently high nor low. She was informed that 10 more days of classroom recording would be needed. Only the first 5 days of the recording constituted Condition 4, using the same procedures as Condition 2, while the next 5 days' recording were used to train raters for the present study.

Debriefing. On completion of recording, the teacher was told the true nature of the study. She was given permission to view all recordings, and it was agreed that data would not be used without her expressed permission. The teacher reported no knowledge of the true nature of the study while it was in progress.

Behavioral ratings. Two male undergraduates[4] were selected to be behavioral raters and received course credit for participation. Given the same "blind" as the teacher, they were told they were the the videotape raters in a study assessing the agreement between an observer in a classroom and a videotape observer. Each was given a manual describing the "blind" study and the behavioral rating forms. Since the presence of the observer in the classroom could not be determined by viewing the tapes, the raters believed that the observer had always been present.

The raters were first trained to rate classroom behaviors with a modified version of the Teacher–Pupil Interaction Chronograph (TPIC; Meyers & Craighead, 1973). Each TPIC form was composed of six 8 × 10 matrices; each matrix was used to rate 30 seconds of interaction between the teacher and pupils, three minutes per form. The following eight categories of pupil behavior were represented along the top of each matrix:

1. Gross motor—getting out of chair, standing up, walking around, moving chairs, running (unless specified by teacher).

2. Orienting—turning head or body to interact with another person not specifically instructed by the teacher, must be of four seconds duration, 45 degrees or more for the body and/or 90 degrees or more for the head, using the chair as a reference.

3. Object noise—tapping tokens or other ob-

---

[4] Appreciation is expressed to David Barnes, Norvin Cooley, and William Fernan, who served as raters. Appreciation is also expressed to Jerome Gordon, Ronald Madle, and Laurent Hahn for their assistance with the videotape aspects of this study.

jects, clapping, tapping of feet (unless specifically instructed by teacher).

4. Disturbance of others and contact (nonverbal)—disturbance of other's property or tokens, throwing objects at another pupil, hitting, pinching, kicking, shoving, slapping, striking with thrown object, poking (any physical contact unless specifically instructed by teacher).

5. Verbalization—inappropriate conversation with another student, answering teacher without being called on, making comments or calling out, calling teacher's name to gain attention, crying, singing, screaming, laughing (they must be inappropriate).

6. Ignoring—ignoring the teacher, may be involved in other inappropriate behavior or may be just sitting doing nothing whatsoever.

7. Other inappropriate—other time off teacher defined tasks not included in categories listed above.

8. Appropriate behavior—time on teacher defined tasks.

Ten categories of teacher behavior were arranged along the left side of the matrix:

1. Positive prompt—a statement of positive expectations or requirements which incorporates a reference to the positive consequences of the behavior for the student.

2. Negative prompt—a statement of negative expectations or requirements which incorporates a reference to the negative consequences of the behavior for the student.

3. Positive statement—expression of expectation, encouragement, a command, instruction or suggestion without reference to the consequences.

4. Negative statement—expression of expectations, discouragement, prohibition, without reference to consequences.

5. Positive verbal feedback—compliments, praise, verbal positive reinforcement, positive feedback, often of the form of repeating correct pupil verbal response.

6. Negative verbal feedback—reprimands, negative feedback, derogatory remarks, audible "shushing."

7. Positive nonverbal feedback—smiling, pleasant expression, positive gesture, affectionate or complimentary pat.

8. Negative nonverbal feedback—frowns, grimaces, negative gestures, also pulling, shoving, or restraining the student.

9. Ignore, no response—behavior of the student followed by no response by teacher, she turns away or does not react verbally or nonverbally, although it must be obvious that the teacher has seen or heard the behavior.

10. Neutral—questions or conversation without positive or negative valence, classroom announcements.

By matching a particular row with a particular column (or column with row) and recording the initial of the pupil with whom the teacher was interacting, the raters could record the teacher–pupil interactions. The first four teacher behaviors had to precede the behavior of a student, and Teacher Behaviors 5 through 9 had to follow the behavior of a student. Neutral teacher behavior could precede or follow student behavior. Each token dispensed during an interaction was recorded by placing a check over the student's initial.

During training the experimenter and the two raters recorded simultaneously, one TPIC form at a time. Training continued until percentage agreements (Number of Rater Agreements on Recorded Behaviors divided by the Number of Recorded Agreements plus Disagreements times 100) per TPIC form between the raters and between each rater and the experimenter were 83% or above for the equivalent of five days of classroom rating, 15 TPIC forms per rater.

Following TPIC training, the taped classes were randomized and assigned to the raters to rate individually. Every fifth class served as a reliability check. The experimenter and one rater were present during independent ratings; the experimenter ran the equipment. The rater began rating at an arbitrarily predetermined time on the tape (ostensibly the time the classroom observer began his ratings) by beginning his stopwatch when that time was visible on a clock in the taped classroom. The rater completed nine minutes of TPIC rating per class period. He viewed the assigned section of tape as many times as he felt necessary for accuracy. The procedure for reliability checks was the same except both raters were present, rating simultaneously.

On completion of TPIC ratings, the raters were trained with the Overall Behavior Checklist (OBC), a modified form of the Time-Sample Behavior Checklist described elsewhere (Meyers & Craighead, 1973). Each OBC yielded four minutes of behavioral ratings for each student regardless of interactions with the teacher. Each minute was divided into six 10-second intervals; five to record behaviors and the sixth to make comments on the checklist. It was possible to record any of eight student behaviors (those listed for the TPIC, above) in each 10-second interval. Training procedures were identical to TPIC training and continued until the percent agreements per OBC between raters and between each rater and the experimenter were consistently above 85% for the equivalent of rating three students per day for five days or 15 OBCs per rater.

Following OBC training, the taped classes were again randomized and assigned to raters. Every fifth class served as a reliability check. Rating procedures were similar to TPIC monitoring, except the time to begin rating was set later in each class to avoid repeating sections covered by TPIC rating. The rater completed four minutes of OBC rating for each student; each student was rated separately over the same time period. The rater viewed the assigned section of tape as many times as necessary for accuracy; the rating was done during the last viewing of each tape.

Following OBC ratings, the raters were trained to count the number of tokens dispensed per minute by using five 10-minute periods, the equivalent of 10 minutes of rating per class for five classes; 100% agreement was obtained for all five 10-minute training trials. The raters were told token counting was necessary to determine the accuracy of the TPIC in assessing the number of tokens dispensed per class.

Because of perfect training agreement only one rater counted tokens for all classes. The rater began counting at the identical times as were used for TPIC ratings and counted for 10 minutes per class.

Dependent variables. Six dependent variables were obtained from the TPIC, OBC, and token count. They were: (*a*) number of teacher–pupil interactions; (*b*) proportion of appropriate teacher–pupil interactions; (*c*) proportion of inappropriate teacher–pupil interactions; (*d*) proportion of neutral teacher–pupil interactions; (*e*) proportion of appropriate student behaviors; and (*f*) number of tokens dispensed.

The number of teacher–pupil interactions was determined by counting the number of interactions recorded on the TPIC per day. The proportions of appropriate, inappropriate, and neutral teacher–pupil interactions were determined by first arbitrarily designating cells in the TPIC matrix as appropriate, inappropriate, and neutral from a behavior modification conceptualization (see Meyers & Craighead, 1973). Postive prompts, positive statements, positive verbal feedback and positive nonverbal feedback occurring in conjunction with appropriate student behavior, and teacher ignoring occurring in conjunction with inappropriate student behaviors were defined as appropriate teacher–pupil interactions. Neutral teacher behavior occurring in conjunction with any student behavior was defined as neutral teacher–pupil interaction. All remaining combinations of behaviors were defined as inappropriate teacher–pupil interactions. The proportion appropriate was then determined by dividing the number of appropriate interactions per day by the total number of interactions per day. The proportions inappropriate and neutral were similarly calculated. The proportion appropriate student behavior was determined by finding the number of appropriate 10-second intervals per pupil per day on the OBC and dividing that by 20, the total possible number of appropriate 10-second intervals per day. The number of tokens dispensed was calculated by counting the number of tokens per 10-minute interval per day.

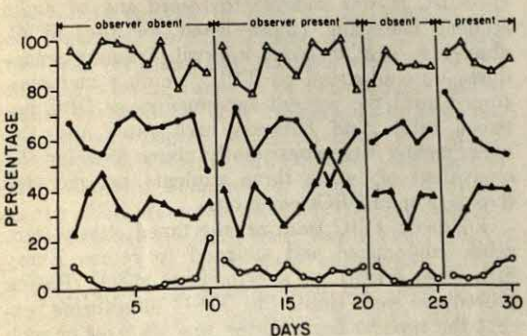


FIGURE 1. The percentage of appropriate student behavior for all students (open triangles) and the percentage of appropriate (solid circles), inappropriate (open circles), and neutral (solid triangles) teacher–pupil interactions for each day of each treatment condition.

RESULTS

Rater percent reliabilities (Number of Rater Agreements on Recorded Behaviors divided by the Number of Recorded Agreements plus Disagreements times 100) for the TPIC and OBC were calculated at reliability checks every fifth taped class. TPIC reliabilities ranged from 84% to 100%, with a mean of 92.4%. OBC reliabilities ranged from 85% to 100%, with a mean of 97.8%. Because of 100% agreement during token count training, no reliability checks were made for the token count.

Figure 1 presents the percentage of appropriate student behavior per day for all students under Conditions 1, 2, 3, and 4. After arc sine transformations for proportional data, these data were submitted to one-way repeated measures analysis of variance for unequal cell frequencies; there were no significant differences in appropriate student behavior across treatment conditions ($F < 1.00$).

The percentage appropriate, inappropriate, and neutral teacher–pupil interactions under Conditions 1 through 4 are also presented in Figure 1. After arc sine transformations for the proportional data, they also were submitted to one-way repeated measures analyses of variance for unequal cell frequencies. No significant treatment effects were found in any of the three analyses ($F < 1.00$; $F = 1.186$, $df = 3/26$; $F < 1.00$, respectively).

Figure 2 presents the number of teacher–pupil interactions per day under Conditions 1, 2, 3, and 4. These data were submitted to one-way repeated measures analysis of variance for unequal cell frequencies, and a significant difference was found between conditions ($F = 6.171$, $df = 3/26$, $p < .003$).

A Scheffé post hoc complex comparison of means for Conditions 1 and 3 versus 2 and 4 was significant ($-39.37 < \psi_g < -3.73$, $p < .05$), with the mean for Conditions 2 and 4 being greater. T tests for differences among means of groups with unequal *N*s were then performed on the data. The means of Conditions 1 and 3 were not significantly different ($t = 1.40$, $df = 26$) nor were the means of Conditions 2 and 4 ($t = .70$, $df = 26$). The mean of Condition 1 was significantly different from the means of Conditions 2 and 4 ($t = 2.91$, $df = 26$, $p < .05$ and $t = 3.61$, $df = 26$, $p < .05$, respectively) with the means of Conditions 2 and 4 being greater. The mean of Condition 2 was not significantly different from the mean of Condition 3 ($t = 1.50$, $df = 26$), although the mean of Condition 3 was less (see Figure 2). Lastly, the mean of Condition 3 was significantly different from the mean of Condition 4 ($t = 2.21$, $df = 26$, $p < .05$), with the mean of Condition 4 being greater.

Figure 2 also presents the number of tokens dispensed per day under Conditions 1, 2, 3, and 4. These data were submitted to one-way repeated measures analysis of variance for unequal cell frequencies; no significant differences ($p < .05$) between conditions were obtained ($F = 2.792$, $df = 3/26$, $p = .06$). A Pearson product-moment correlational analysis yielded a significant positive correlation ($r = .74$, $df = 58$, $p < .005$) between the number of tokens dispensed and number of teacher–pupil interactions per day.

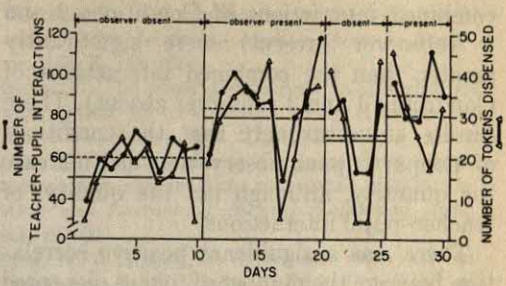


FIGURE 2. The total number of teacher–pupil interactions (solid circles, condition means indicated by broken lines) and the total number of tokens dispensed by the teacher (open triangles, condition means indicated by solid lines) for each day of each treatment condition.

DISCUSSION

OBC data revealed no significant differences in the percentages of appropriate student behavior across conditions. The initial high level of appropriate student behavior may have precluded the demonstration of an experimental effect in the direction of more appropriate behavior during observation. The quality of teacher–pupil interactions, assessed by the TPIC, was not affected by observation; neither the percentage of appropriate, inappropriate, nor neutral teacher–pupil interactions changed significantly across conditions. The low initial baseline made it difficult to demonstrate a decrease in percentage of inappropriate behavior as a result of observation, such as was suggested by White's (1973) data.

The TPIC data did indicate a significant effect on the quantity of teacher–pupil interactions. When nonparticipant observation was instituted (Condition 2), the number of interactions increased significantly; when removed (Condition 3), the number of interactions decreased with the reversal approaching statistical significance. With the reinstatement of observation (Condition 4) the number of interactions again increased significantly.

If reversibility does not occur in the ABAB design, the effect of the experimental variable remains unclear. In the present study the reversal only approached significance. If the effects of treatment conditions are not completely transient and reversible, the behavior produced in the experimental condition may show some resistance to extinction when there is a return to baseline conditions (Kazdin, 1973). Perhaps increased classroom interaction is not a transient, readily reversible phenomena. Anecdotally, the teacher reported increased familiarity with the class during the study. If this increased familiarity resulted in increased interaction over the period of the study, beyond the effects of the observational process, the obtained results are not surprising.

It is clear, however, that significant increases in interaction were associated with the introduction of observation. Further, the

combined interactions of Conditions 2 and 4 (observer present) were significantly greater than the combined interactions of Conditions 1 and 3 (observer absent). These results argue strongly that the conditions of nonparticipant observation did increase the quantity, although not the quality, of teacher–pupil interactions.

There was a significant positive correlation between the number of tokens dispensed per day and the number of interactions per day. Further, the trend in the mean number of tokens dispensed per condition was similar to that of the mean number of interactions per condition. It is difficult to determine whether the number of tokens varied as a function of the number of interactions, vice versa, or if both varied as a function of some third variable (e.g., presence of observation).

The data do not indicate habituation to observation; in fact, the teacher appeared periodically resensitized to observation as indicated by the increased number of interactions approximately half way through each observational condition. It has been hypothesized that at least 10 to 12 days are required for habituation to an observer's presence (Medley & Mitzel, 1963). If this is true, it is highly unlikely that effective habituation would have occurred during the present study's observational conditions, 10 and 5 days respectively. Even if habituation is a demonstrable phenomenon, it may not be feasible economically to have observers present long enough for the habituation to take place; moreover, the researcher's choice of the point at which the subjects' behavior is "normal" (i.e., free from observer influence) is an arbitrary one (Masling & Stern, 1969).

Other observational reactivity was noted anecdotally. The teacher related an "inability" to act normally while being observed and asked the experimenter how he would like her to act for the study (she was told to act normally). Like Gussow's (1964) teachers, she showed interest in what was being recorded and in what manner the records were to be used. Although data were not recorded (at the teacher's request), it appeared to the experimenter that the amount of time standing in front of the classroom while teaching varied as a function of conditions. The students attempted to interact with the experimenter as he observed in the classroom and occasionally asked the experimenter after class what he had been recording. They were also overheard telling other residents of the institution they were "on TV." Nevertheless, student cognizance of observation apparently had no effect on student behavior across conditions.

The effect of increased interactions cannot be attributed solely to the presence of an observer. Under Conditions 2 and 4 (observer present) the subjects were aware of being recorded on videotape for the "blind study." The systematic confounding of videotape recording with observer presence made the independent variable nonparticipant observation. A cleaner design would call for the use of concealed observation undetectable by the subjects.

The wide range of student ages and years of institutionalization and their more limited range of IQs, not atypical of the institutionalized retardate, argue for generalization to a large female resident population. The generalization from the sample teacher is perhaps more limited. She appeared very capable in applying behavior modification principles to classroom control (e.g., low levels of inappropriate teacher–pupil interactions under all conditions). A competent teacher may have little reason to feel threatened by the presence of an observer familiar with the principles applied to her teaching and therefore be unlikely to change her overt behavior. Further, both the teacher and the pupils were observed on a periodic basis within the institution, though not in this particular class. A greater reactivity may be detected in a setting where periodic observation is not the rule.

Feedback from observers concerning attending behavior to appropriate child responses can alter teacher attending behavior (Cooper, Thomson, & Baer, 1970) and data (Panyon, Boozer, & Morris, 1970) have suggested that feedback to attendants may act as a reinforcer for applying operant techniques. It has also been noted in reinforcement programs where the presence of

an observer is associated with implementation of a program to establish behavior that there is the possibility of the target behaviors coming under discriminative control of the presence of the observer (Kazdin, 1973; Surratt, Ulrich, & Hawkins, 1969). Since the present study associated neither feedback nor the application of operant techniques with observation, generalization to studies which have included such procedures should be drawn cautiously. When feedback is involved, the potential for greater reactivity exists. However, since the speed of acquisition of a target behavior is a direct function of the number of reinforcements, one must also be cautious in comparing the findings of studies which differ in whether an observer was present or absent; the present findings suggest that the presence of an observer increases the number of interactions, and thus the frequency of reinforcement.

The finding that observation was associated with increased frequency of teacher–pupil interaction would not have been predicted from White's (1973) finding that observer presence was associated with decreased activity level of mother–child dyads. These opposite findings suggest that the nature of observer reactivity may be situation specific; therefore, it is essential that the phenomenon be evaluated in a variety of settings. It may be concluded that observational reactivity does occur, however minimal its effects, and it should be taken into account in generalizing from observational studies. Further research should indicate the parameters which are important contributors to the "observation reactivity" limitations on generalization.

REFERENCES

Baer, D. M., Wolf, M. M., & Risley, T. R. Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1968, **1**, 91–97.

Cooper, M. L., Thomson, C. L., & Baer, D. M. The experimental modification of teacher attending behavior. *Journal of Applied Behavior Analysis*, 1970, **3**, 153–157.

Gussow, Z. The observer-observed relationship as information about structure in small group research. *Psychiatry*, 1964, **27**, 230–247.

Heber, R. Modifications in the manual on terminology and classification in mental retardation. *American Journal of Mental Deficiency*, 1961, **65**, 499–500.

Heyns, R., & Lippitt, R. Systematic observational techniques. In G. Lindzey (Ed.), *Handbook of social psychology*. Vol. 1. Cambridge, Mass.: Addison-Wesley, 1954.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. Paper presented at the Fourth Banff International Conference on Behavior Modification, Banff, Canada, March 1972.

Kazdin, A. E. Methodological and assessment considerations in evaluating reinforcement programs in applied settings. *Journal of Applied Behavior Analysis* 1973, **6**, 517–539.

Masling, J., & Stern, G. Effect of the observer in the classroom. *Journal of Educational Psychology*, 1969, **60**, 351–354.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Meyers, A. W., & Craighead, W. E. Behavior modification with the autistic child in the classroom. Paper presented at the meeting of the Association for the Advancement of Behavior Therapy, Miami, Fla, December 1973.

O'Leary, K. D., & O'Leary, S. (Eds.) *Classroom management: The successful use of behavior modification*. New York: Pergamon Press, 1972.

Panyon, M., Boozer, H., & Morris, N. Feedback to attendants as a reinforcer for applying operant techniques. *Journal of Applied Behavior Analysis*, 1970, **3**, 1–4.

Surratt, P. R., Ulrich, R. E., & Hawkins, R. P. An elementary student as a behavior engineer. *Journal of Applied Behavior Analysis*, 1969, **2**, 65–72.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. *Unobtrusive measures: Nonreactive research in social sciences*. Chicago: Rand McNally, 1966.

White, G. The effects of observer presence on mother and child behavior. Paper presented at the meeting of the Western Psychological Association, Anaheim, Calif., April 1973.

(Received August 14, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 520-529

ORIGINS OF TEACHERS' ATTITUDES TOWARD YOUNG CHILDREN[1]

SHERRY WILLIS[2]
The Pennsylvania State University

JERE BROPHY
University of Texas at Austin

Several previous investigations have revealed contrasting patterns of interaction by teachers with students toward whom they hold attitudes of attachment, concern, indifference, or rejection. The present study focused on discovering the origins of these teacher attitudes by collecting first-grade teachers' early impressions of students. Teachers were interviewed during the first two weeks of school, prior to readiness test administration, and their comments about students whom they later assigned to each of the four attitude groups were analyzed. The findings were quite consistent in yielding distinct profiles for children in each respective group, complementing and extending previous research by indicating the student attributes associated with the formation of these four teacher attitudes.

Silberman (1969) asked teachers to nominate one student in their class to each of four attitude groups: (*a*) *attachment* (If you could keep one student another year for the sheer joy of it, whom would you pick?); (*b*) *indifference* (If a parent were to drop in unannounced for a conference, whose child would you be least prepared to talk about?); (*c*) *concern* (If you could devote all your attention to a child who concerned you a great deal, whom would you pick?); (*d*) *rejection* (If your class was to be reduced by one child, whom would you be relieved to have removed?).

He than observed for 20 hours in each class to discover what these children were like and how their teachers interacted with them. He found the attachment students to be "model" students, high achievers who conformed to the teachers' wishes and fulfilled their personal needs. Observation of the teachers interacting with these attachment students showed some evidence of subtle favoritism, but no gross favoritism.

Concern students tended to be dependent, low-achieving students who made extensive but approved and appropriate demands upon the teachers. The teachers interacted most frequently with these students and in general behaved in ways consonant with their expressed concern about these students' achievement levels.

The indifference students did not have any particular identifying characteristics except for their low frequencies of interaction with the teachers. In observing the teachers, Silberman noted that teachers' interactions with indifference students not only were infrequent but also were briefer and less emotionally involving than those with other students.

The rejection students tended to be behavior problems who made demands that the teachers saw as illegitimate or overwhelming. Teachers had frequent contacts with these students, but a large proportion of these contacts involved intervention to control their misbehavior. Yet these students also received considerable teacher praise, as if the teachers were attempting to "make up for" the generally negative tenor of their interaction with them.

[1] The authors wish to acknowledge and thank the teachers and administrators of the Memphis and Shelby County, Tennessee Public School Systems, who made the data-gathering phase of this project possible; Anita Johnson and W. G. Duvall, who served as interviewers for the study; and Jean Waltman, who assisted in manuscript preparation.

[2] Requests for reprints should be sent to Sherry L. Willis, Division of Individual and Family Studies, College of Human Development, The Pennsylvania State University, University Park, Pennsylvania 16802.

Additional research on the student characteristics and teacher–student interaction patterns involving students in Silberman's four attitude groups has been done by Jenkins (1972), Good and Brophy (1972), and Brophy and Good (1974). All of these studies generally support Silberman's results and impressions, although there are some inconsistencies. The teachers in Jenkins' study saw the attachment students as high achievers and as students with warm rather than neutral or hostile attitudes toward themselves (the teachers). Good and Brophy (1972) found the attachment students to be well-behaved high achievers. Similar findings were obtained by Brophy and Good (1974). Thus, in general, attachment students appear to be high-achieving students who conform and respond warmly to teachers. Nevertheless, none of the four studies found any grossly overt favoritism of these students, although they all suggested that the teachers favored attachment students in subtle ways that were not being picked up by the behavioral measures included.

The four studies agree in showing the concern students to be low achievers (the primary reason for teacher concern) and also suggest that these students are dependent upon teachers but express this dependency in ways that the teachers find rewarding or at least acceptable. Thus, they respond to these students with concern about their low achievement and with redoubled efforts to do something about it through more frequent contacts and provision of tutorial help and other aids to learning.

The four studies find that rejection students are mostly low achievers, but not as consistently low as the concern students. However, students in the rejection group apparently "turn off" teachers. This may be done actively through defiance or disobedience, or more subtly through failure to respond positively to the teachers' overtures. Findings regarding teacher interactions with rejection students are mixed. Most studies have found the pattern noticed by Silberman: evidence of a strained teacher–student relationship featuring frequent disciplinary contacts and criticism but tempered by frequent praise and other evidence of teacher attempts to "make up for" negative behavior. However, Good and Brophy (1972) found a clear-cut pattern of rejection untempered by any evidence of attempts to compensate by providing positive behavior toward rejection students.

A common thread running throughout these studies is that students in the concern and rejection groups appear to be quite similar at the gross observational level, with the exception that rejection-group students are more likely to be classroom discipline problems than concern-group students. Teachers show sharply contrasting reactions to these two types of students, however. They tend to be quite supportive and attentive primarily to the *academic* needs of students in the concern group but to be rejecting and primarily attentive to the *conduct* of students in the rejection group.

The four studies agree in finding students in the indifference group to be characterized primarily by low rates of interaction with the teacher. Likewise, data from some of the studies support Silberman's observations that teacher interactions with them are brief and low in emotional intensity. Whereas teachers tend to respond warmly to the attachment group, become concerned about students in the concern group, and reject or develop conflictual responses toward students in the rejection group, they respond with indifference or apathy toward students in the indifference group, often acting as if they were unaware that these students were even in the room. The student attributes that trigger this reaction (or, more properly, nonreaction) have not been identified, however.

The purpose of the present study was to further explore these four attitude groups to try to identify the student characteristics that trigger these four attitudinal responses in teachers. More particularly, this study sought to identify some of the descriptive characteristics of indifference-group students and some of the differences between concern- and rejection-group students which might explain the strongly contrasting teacher reactions to these two groups.

Method

Subjects

Subjects were 28 female first-grade teachers and their students, drawn from a large urban public school system. The students were almost all from white, middle-class homes. Due to a teacher cross-over plan for desegregation purposes, 9 of the teachers were black. All teachers worked in self-contained classrooms in which they taught their children all of the basic academic subjects. Teachers working in team-teaching arrangements or in special education classes were not included in the study. The school system was selected because its children do not attend kindergarten (unless they attend private kindergartens). Thus, when the children entered the first-grade classrooms of the teachers in the study, they were unknown to the teachers and did not bring with them records containing test data or information about kindergarten performance. Therefore, except in a few cases where the teachers were neighbors of the children or where they knew the family from having taught an older sibling, the teachers and students were unfamiliar with one another before school began.

Procedure

Each teacher was seen four times, once to explain the purpose of the study and solicit her cooperation, and three times to conduct interviews in her classroom after school hours. At the first meeting, teachers were informed that they would be interviewed at three points during the year to find out what they had noticed about each of their students. It was explained that very little research information is available about the kinds of student characteristics that teachers notice and use in forming impressions about students and that the purpose of the study was to gather such information. Teachers were informed that interviews would be informal but would be tape recorded and were assured that all information would be strictly confidential. No difficulty was encountered in obtaining cooperation from teachers. They found participation in the study to be stimulating and enjoyable, and the guaranteed confidentiality apparently eliminated any hesitation that they may have had about participating or about speaking freely.

Interviews were conducted by the first author and by two middle-aged female assistants trained by her for this purpose. Interviews were conducted at three points: (*a*) during the first two weeks of school, before any test data were available; (*b*) one to two weeks after the Metropolitan Readiness Test had been administered and scored by the teachers (about four weeks after the beginning of school); and (*c*) during the second and third weeks in January, at the beginning of the second semester. Interviews averaged about an hour, although they varied by teacher. For the first interview, the teacher received the following instructions:

> The purpose of this interview is for you to discuss briefly each child in your class. In discussing each child, you should indicate the characteristics and actions you have noticed about each child. Do not try to limit what you say to only one type of information—just mention anything you know or have observed about this child. You don't have to use complete sentences —you may just give some adjectives or phrases which describe the child.

Directions for the second and third interviews were as follows:

> We will be discussing your students as we did the last time. Now that you have had your students for approximately a month [four months at the third interview], we are interested in what you have noticed about them. Feel free to repeat anything you said last time about a child or to add anything new you have noticed about the child. You don't have to use complete sentences —you may just give some adjectives or phrases which describe the child.

Following these instructions, the interviewer then began naming the children in the teacher's class and asked her to respond, following a prearranged randomized order of names. Once the interview began, the interviewer's primary function was to act as an interested listener and to encourage the teacher to talk comfortably. Since the interviews were tape recorded, the interviewers did not have to take down the information and could spend their time responding to the teacher and encouraging her to continue talking. Ground rules were set up for interviewers to help insure that they did not cue teachers to give certain kinds of information or begin to reinforce them for certain kinds of comments. Also, teachers were instructed to confine their responses to what they had observed about the child and to omit discussions of test scores, attendance data, or long anecdotes which were intended to exemplify or justify a general statement about the child that had just been made. The interviewer avoided making evaluative comments on the teacher's statements. Teachers were encouraged to continue talking about a given child as long as they had additional things to say. When they seemed to have exhausted their perceptions about a child, the interviewer then named the next child.

After each interview, the teachers were asked to rank their students in order according to their expected achievement levels. These data were used as part of an investigation of the student characteristics noticed and used by teachers in forming impressions about students' academic potential (see Willis, 1972). Also, following the third interview, the teachers were asked to nominate up to three students to each of the four attitude groups (attachment, indifference, concern, and rejection), using the method of Silberman (1969) mentioned previously. These attitude-group nominations formed the basis for the present study, since they allowed investigation of the characteristics of students nominated to these four attitude groups.

Coding the Interviews

The tape-recorded interviews were transcribed and then coded according to a system devised by the present authors. Coding categories had not been established prior to the interviews; they were established after the interviews had been completed, based upon observed responses. An initial system was devised in response to a set of five interviews, and the system was then revised after it had been tried out with an additional five interviews. The revised system was then used to code all of the interviews, with intercoder agreement being 85% (number agreed divided by itself plus number disagreed). Disagreements were resolved by discussion. During initial coding and resolution of coding differences, the authors did not know whether or not the child in question had been nominated to one of the four attitude groups by his teacher, so that the results of the study could not have been biased by the authors' knowledge of the teacher's general attitude toward the student.

The teachers' interview responses about each student were classified into 11 major divisions. There were 117 specific categories within these 11 major divisions, as well as a general/unclassified category for each division to handle statements which fell into that division but which did not clearly fit into one of the more molecular categories. The 11 divisions included statements concerning the following: the physical description of the child, the child's family, the child's health or physical condition, the child's social or emotional characteristics, the child's interaction with other children, the child's attitudes or motivation concerning school, the child's classroom behavior or the kinds of management problems that he presented, the child's readiness for school or readiness-related abilities and skills, the child's oral or verbal skills, the child's work habits or ability to do classroom work, and the teacher's feelings concerning the child, the nature of their relationship, or individual interactions that she had had with the child.

In addition to coding each statement within one of the categories in the system, the statement was coded for whether it was positive, neutral, or negative. Thus, the coding yielded information both about what kinds of things the teacher had noticed about the child and about whether the information was positive, neutral, or negative.

Data for the present study were taken from the first interview, done during the first two weeks of school. Thus, they reflect the teachers' early impressions of the students, based solely on contact with them in or out of class. No standardized test data were available yet.

Data Analysis

Each child was scored present or absent for each of the basic categories in the system. Thus, he had a presence/absence score for each of 117 specific categories and 11 unclassified/general categories. One hundred and twenty additional scores were derived through algebraic transformation of these original 128 categories, yielding 248 scores. The 120 additional scores were sum scores for each of the 11 major categories and sums of positive and negative statements within and across categories.

To assess the relationship between perceived characteristics of the students and assignment of students to the four attitude groups, a series of one-way analyses of variance was performed, in which students nominated to a given attitude group were compared with all other students of the same sex. Comparisons were made separately within sex, because there is evidence that attributes associated with sex role expectations are perceived more positively when they appear in a student of the "appropriate" sex than when they appear in a student of the opposite sex (Brophy & Good, 1974; Feshbach, 1969). All students were included in the analyses except those who were repeating the first grade. The latter were excluded because teachers had known them for an extra year and because their status as repeaters probably affected the teachers' attitudes toward them in systematic but unknown ways.

These 248 comparisons for each attitude group involved 355 boys and 311 girls. Given the large *N*s, it might seem that statistically significant *F* values could be obtained even where group differences were minor in absolute magnitude. However, with the exception of the sum scores for the major categories and a few heavily used categories, incidence of category use was low enough to prevent many differences from reaching the .05 level of significance. In fact, category usage was often so low that a meaningful analysis could not be performed. Thus, the .05 level of statistical significance was chosen as the criterion for inclusion of a group-difference finding in this report; all group differences reported below are based on *F* values at or below the .05 significance level.

Results

The distributions of boys and girls in the four groups are shown in Table 1. Although the differences are not large (except in the

TABLE 1

Distributions of Boys and Girls in the Four Attitude Groups

Distribution	Group							
	Attachment		Concern		Indifference		Rejection	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
n	32	32	42	22	30	23	37	25
%[a]	9.7	11.2	12.7	7.7	9.1	8.0	11.2	8.7

[a] Percentage of total number of boys and girls, respectively, who were nominated to each attitude group (each teacher named as many as three children to each group).

concern group), they reaffirm the frequent findings that boys are more salient than girls (more likely to be noticed and commented upon) and more likely to be perceived negatively by teachers (Brophy & Good, 1974).

As expected, teachers' comments about children in the *attachment group* were overwhelmingly positive. Concerning the *boys* nominated to the attachment group, the teachers made more positive comments about their clothing, more often said that they had an immature appearance, more often said that they had a visual impairment or required glasses, less often said that they were quiet, more often assigned them as leaders or classroom helpers, more often described them as helpful with other children, more often described them as busybodies, more often stated that they knew left from right and could stay within lines on a tablet (readiness skills), more often stated that they did not draw well, more often made negative comments about their reading ability, more often stated that they volunteered information during classroom discussions, more often mentioned a perceptual problem or learning disability, more often mentioned positive classroom behavior, more often mentioned positive social behavior, and more often mentioned the student as a high-ability student. The latter perceptions were confirmed by these students' significantly higher scores on the Metropolitan Readiness Test. In general, boys in the attachment group appeared to be high-ability students who were well adjusted to school, conformed to the teachers' rules, and "rewarded" the teachers by being supportive of them by doing well in their school work, helping out, and volunteering information. The negative statements regarding student ability and reading progress were probably relative rather than absolute, in view of the more general picture of high ability that the teachers drew in describing these boys. In the context of their total statements about these boys, these negative remarks appear to be more a matter of concern about getting the boy to maximize his potential rather than concern about getting him to meet minimal requirements. The comments about clothing suggest that these boys were generally of higher socioeconomic status than their classmates. Teacher favoritism toward attachment boys is suggested by at least one set of findings: even though these boys were perceived as busybodies, they were more often assigned as leaders or helpers than other boys were.

Teachers described *girls* assigned to the *attachment group* as larger than average, more attractive than average, having interested, cooperative parents, more likely to have a visual impairment or require glasses, more likely to have been to kindergarten, more likely to be unable to write their names, more creative and imaginative, more alert, better observers, more likely to enjoy stories, able to work independently on assignments, having higher general intellectual ability, coming from generally good families, and being high in expected achievement. Just as with the boys, the teachers' statements about high ability in the attachment group girls were borne out by significantly higher Metropolitan Readiness Test scores. In addition to these specifics, the teachers had significantly more positive comments and significantly fewer negative comments about these girls than about other girls. Like the boys, the girls in the attachment group appeared to be high-achievement students who conformed to and rewarded their teachers. Also, there again is evidence that they came from higher-social-class families and that they were perceived in a generally positive way, including physical attractiveness.

These data generally parallel previous findings suggesting that teachers hold a general positive halo effect toward children assigned to the attachment groups and that, even when negative qualities are perceived in these students, the teachers do not respond to them negatively as they apparently do in responding to similar traits in students in the other three groups.

Boys assigned to the *concern group* were especially likely to be described as being of average size, being reared by grandparents or older parents, having a speech impediment or using baby talk, being generally immature, being active and vivacious, seeking teacher attention, being able to use and

keep up with school supplies, being dependent in school work and needing help from the teacher, needing reassurance and approval, being of generally low ability, needing readiness work, having a positive attitude toward school, having generally poor health, having generally poor social–emotional development, having generally poor oral and verbal skills, having generally poor skills in the area of independent work, and having generally low abilities. The latter perception was confirmed by significantly lower Metropolitan Readiness Test scores for these boys. In addition, the teachers made significantly fewer positive comments and significantly more negative comments about these boys, although the negative comments were almost completely confined to their abilities rather than to their personalities or cooperation with the teachers.

Taken together, these data suggest a rather clear picture of the boys in the concern group. They appear to be students of low ability who are dependent upon the teachers for help and reassurance in completing their assignments. However, they are also perceived as being cooperative and compliant, so that demands they make upon the teachers are perceived as being legitimate. The teachers also perceive them as poorly adjusted in their social relationships and generally immature, again stressing their dependency upon and need for help from the teachers.

Teachers describe *girls* in the *concern group* as being more likely to be nonwhite than to be white; they mentioned the parents' occupation more frequently (often with a negative connotation); their parents were described in generally positive terms; they tended to be from large families; the size of the family was one of the problems that the teachers felt was confronting the child; the child was more likely to have a speech impediment or use baby talk, to be more dependent and quieter, to be lacking in self-confidence, to be dependent in work, and to need help from the teacher, to have a generally positive attitude toward school, and to have poorly developed verbal skills. Again, these girls had significantly lower Metropolitan Readiness Test scores than their classmates.

Even more than the boys, the girls in the concern group showed a pattern of low achievement combined with dependency upon the teacher for both emotional and academic support. The data for both boys and girls in the concern group in this study suggest that concern students are low achievers who depend on the teacher to tell them what to do continually. This dependency is expressed in inhibited, socially approved ways; concern children apparently do not present behavioral problems and are not assertive or intrusive like children in some of the other groups.

Teachers describe *boys* in the *indifference group* as more likely to be blond-haired, to have a "blank" facial expression, to be immature, to be neat and clean, to have a working mother (which the teacher perceives as a problem), to be reared by grandparents or substitute parents, to have a disinterested or uncooperative parent, to have a visual impairment or need glasses, to have a speech impediment and/or speak in baby talk, to be a sociometric loner, to be anxious to please, to be in poor health, to have negative attitudes toward school, to have failed to live up to the teachers' initial expectations, and to have poor verbal skills. Also, in summative response categories about indifference boys, the teachers more often mentioned physical descriptions, more often commented negatively upon their health, more often made negative comments about their work-related behavior, and made more total negative comments about them. Despite this rather consistently negative pattern, which included several negative comments about abilities and work habits, the boys in the indifference group did not differ significantly from other boys on the Metropolitan Readiness Test.

Some of the contrasts between concern boys and the indifference boys are striking. Both groups were perceived as being of low ability (especially the concern boys), but the two groups of boys were perceived as responding to teachers in contrasting ways. Whereas the concern boys showed dependency upon the teachers, the indifference boys apparently did not respond to the

teacher in ways that were rewarding to them. Even though the teachers perceived the needs of these children quite clearly and even though they rated them as anxious to please, they responded to them with indifference rather than with concern. Clues indicating possible reasons for this appear in some of the other teacher perceptions. In particular, the teachers reported that indifference boys had a "blank" facial expression (suggesting that they did not respond to teacher overtures) and that they had poor attitudes toward school (the indifference group was the only group to be so described). They were also described as having failed to meet the teachers' initial expectations for them. Apparently, on the basis of their appearance and of observations of their early work, the teachers had developed moderate to high expectations for these boys. Apparently, however, they responded inappropriately (from the teachers' point of view, at least) to teacher overtures (blank expression, "negative" attitudes), thus conditioning the teachers to stay away from them.

The latter data are reminiscent of the findings of Yarrow, Waxler, and Scott (1971), who found that teachers returned more quickly and more frequently to a child who had given them a positive response to their previous overture than to a child who had failed to give such a positive response. Positive responses to teacher overtures are apparently experienced as rewarding by the teachers, so that children can condition teacher approach or avoidance by either producing or failing to produce, respectively, positive responses to these overtures. Apparently this phenomenon was operating in these classrooms, and it must have exerted a strong primacy effect in conditioning the teachers to avoid and become indifferent toward these boys, even though they clearly perceived their needs for help in certain areas. The teachers apparently felt that these boys wanted to be left alone and/or that they disliked the teachers or school, so they began to avoid them.

The *girls* in the *indifference group* were described as more likely to be nonwhite than white, as not liking school, as easily giving up when working on assignments, as lacking in self-confidence, as not being prepared for school, as not knowing their colors or numbers, as being creative and imaginative, as being of generally low ability, as being children that the teachers did not frequently interact with, and as presenting problems in their classroom behavior. Teacher perceptions of deficient readiness skills in these girls were borne out by the girls' significantly low Metropolitan Readiness Test scores. Here again, the teachers clearly perceived that these children were in need of help, but they responded with indifference rather than with concern.

Again, the apparent reason is the response of the children to the teacher. Whereas the concern group combines low ability with dependency and positive response to the teacher, the indifference group combines low ability with misbehavior in the classroom and negative attitudes toward school. The major difference between the two groups seems to be a negative response to the teacher and/or to school in general among the indifference students. When teachers encounter such attitudes and do not succeed in changing them, they tend to respond with indifference of their own. Apparently, this is not a benign indifference which occurs just because a child is nonsalient and the teacher is too busy interacting with classmates to notice him. Instead, indifference of this sort seems to be a defense mechanism to protect the teacher from continued frustration and rejection by the indifference-group students. Such unresponsive and/or sullen behavior by students during early interactions with the teachers tends to "turn off" the teachers, conditioning them to minimize their interactions with these students in the future and to develop an attitude of indifference rather than concern about their problems, even though these problems are accurately perceived.

Boys in the *rejection group* were described as more likely to be nonwhite than white, as coming from intact families in which both parents were living, as immature and poorly adjusted, as independent,

as loud or disruptive, as being relatively inactive and not vivacious, as unlikely to be assigned as a leader or helper, as having difficulty in getting along with others, as being talkative, as not knowing likenesses and differences, as not knowing how to write their names, as not knowing left from right or how to stay within lines on tablets (readiness skills), as unable to use or keep up with school supplies, as having poor reading abilities, as needing extra help because of generally low ability, as being likely to fail or to have to be withdrawn from school, as having deteriorated in their work since the beginning of the year, as being either notably healthy or having notably poor health, as having poor verbal skills, as being physically unattractive, as presenting classroom behavior problems, as generally lacking in school readiness, as presenting problems in their behavior during work assignments, and as being of generally low ability. Although the teachers remembered both more positive and more negative contacts with these children, they had more total negative comments about them, fewer total positive comments about them, and more total comments about them (positive plus negative). Despite this extremely detailed and almost unremittingly negative picture of the rejection-group boys, including several explicit statements to the effect that they were of low ability, these boys did *not* differ significantly from other boys in Metropolitan Readiness Test Scores.

Just as teacher reactions to the attachment-group students suggest a generally positive halo effect, the reactions to the rejection-group boys suggest a generally negative one. The halo effect induced by their intense dislike of the rejection-group boys even caused the teachers to misjudge (underestimate) seriously their intellectual abilities, even though these teachers generally were remarkably accurate in judging student potential (Willis, 1972).

The contrast with students in other groups is also instructive. Instead of being quiet and dependent upon the teacher like the concern students, the rejection boys are assertive and often loud and disruptive; and they present frequent classroom discipline problems. Their interpersonal problems are not confined to their interactions with the teachers, as with indifference students; they are also described as having difficulty in getting along with their classmates. Thus, the problems that these boys present to the teachers are frequent and serious enough to cause the teachers to respond with rejection rather than with concern. In this connection, it is noteworthy that, although both the concern and the rejection boys are described as in need of readiness work and extra help in general, only the rejection boys are described as likely to fail or to have to be withdrawn from school.

This implies that the teachers had positive expectations of success in their efforts to work with concern students but did not have such expectations for whatever efforts they may have been making to give help to rejection students. Also, it should be kept in mind that concern students had generally low abilities as measured by the readiness tests, while the rejection boys did not differ significantly from other boys. Apparently, boys who present sufficiently severe discipline problems create such a strong negative halo effect that teachers also attribute low abilities and lack of readiness to them, despite evidence to the contrary. This is the only group in which teacher perceptions of ability did not match readiness test data. Apparently, the frustration and aggravation caused by these boys was sufficient to impair teacher judgment in their case. Thus, they underestimated these boys' intellectual abilities and were pessimistic regarding their chances for academic success, even though they were optimistic about the academic success of concern-group boys who actually did have lower abilities.

Concerning the *girls* in the *rejection group*, the teachers frequently mentioned negative family patterns (broken home or poor parental cooperation), and they described the girls as being busybodies, not liking school, giving up easily, lacking self-confidence, being playful and mischievous, not being prepared for school, not being adjusted to school routines, not knowing

colors and numbers, not knowing likenesses and differences, being alert and close observers, not volunteering information to the class, failing to pay good attention, being likely to fail or to have to be withdrawn from school, being able to achieve more than they were doing, having poor school attitudes, having poor school readiness, having poor work habits, and having poor general ability. Again, the teachers made significantly more negative comments, significantly fewer positive comments, and significantly more total comments about the rejection girls than about girls in the other groups. As with the boys, however, despite the teachers' repeated and detailed comments about low abilities, the rejection girls did not differ significantly from their classmates on Metropolitan Readiness Test scores.

Although it shows up in somewhat different variables, the general pattern concerning rejection-group girls is quite similar to, and has the same implications, as the pattern for boys. In contrast to other girls, the rejection-group girls appeared to present more behavioral and disciplinary problems, to be underachievers, and in general not to "go along with the program." As with the boys, the teachers tended to rate the rejection-group girls as having low abilities (although they sometimes described them as underachievers rather than as low-ability children), even though they did not differ significantly from other girls in readiness test scores. The contrast between concern-group girls and rejection-group girls parallels the same contrast for the boys. Concern girls have low abilities but rejection girls do not, although teachers *think* that they do. The teachers apparently were rewarded by the concern-group girls' behavior and developed concern and the habit of spending time with them, but they were apparently put off by the behavior of the rejection students and developed an attitude of rejection toward them.

Discussion

Teacher reactions to the four types of students studied in this research are readily explainable on the basis of the behavior of the students themselves (as perceived by the teachers). The three major variables involved seem to be the students' general level of school success, the degree to which they reward teachers in their personal contacts with them, and the degree to which they conform to classroom rules. *Attachment students* were compliant and successful in school, and they apparently rewarded teachers in their interactions with them. *Concern students* had difficulty in school but apparently were compliant and personally rewarding to the teachers, so that teachers became concerned about them and spent much time providing remedial help. The teachers' negative attitudes toward indifference and rejection children led them to underestimate their ability and learning potential. However, they did not respond with concern even when they accurately perceived the needs of these students, apparently because they were "turned off" by the students' personalities and behavior.

The *indifference students* apparently responded negatively to the teachers, failing to provide a rewarding interpersonal contact pattern, so that the teachers became indifferent and gradually spent less and less time with these children, even though they perceived them as needing extra help. The *rejection students* not only failed to provide rewarding experiences to the teachers in their interpersonal contacts with them; they also frequently caused classroom disturbances and were general discipline problems. The teachers responded to this by rejecting the students to the point of wanting to get rid of them and by projecting a number of traits onto them, especially low abilities, which they did not, in fact, possess (at least not as a group).

Teacher perceptions were generally accurate for the first three groups, but the children of the rejection group apparently were sufficiently threatening to the teachers to impair the accuracy of their perceptions and to cause them to project inappropriate and incorrect attributes onto them. This extended even to teacher judgments of student ability, which were usually quite accurate.

All in all, the data of this study fit nicely

with the findings from Feshbach's (1969) study of teacher preferences for different types of students and from previous studies which related teacher assignment of students to one of these four attitude groups to teacher–student interaction patterns (Brophy & Good, 1974; Good & Brophy, 1972; Jenkins, 1972; Silberman, 1969). They extend this line of research by providing more information about the personal qualities of the students assigned to these four groups.

The data on concern and rejection students are particularly instructive, because previous research had found them to be quite similar, raising the question of why teachers responded to one group with concern while responding to the other group with rejection. The present findings suggest that these two groups are in fact different, although their differences have not shown up clearly in the student behavior measures included within previous observational research. Previous interaction analysis research had identified compliance versus misbehavior as one apparent difference between concern- and rejection-group students, although the findings did not suggest anything resembling the extreme difference suggested by the present findings. Perhaps actually existing extreme differences were not picked up in previous research for some reason, or perhaps teachers in the present study were exaggerating the actual differences as a defense mechanism. Probably both factors were operating to some degree.

The present study has also provided some positive descriptions of the qualities of the indifference-group students. Previous research had identified only low frequency of contact with the teacher as a predictable characteristic of this group. The present research has identified several additional characteristics which help form an explanation for the low frequencies of teacher contact with indifference-group students. The findings suggest that a measure of the degree to which these students provide teachers with rewarding responses in their interpersonal contacts with them (as in the study by Yarrow, Waxler, & Scott, 1971) would show a significant difference between the students in this group and their classmates.

Across attitude groups, a major conclusion of this research is that the particular relationship between a teacher and an individual student is crucial in affecting the teacher's attitudes toward that student, independent of such general student characteristics as achievement, race, sex, etc. It appears that children who do not reward teachers are avoided and/or rejected by them. Also, although some relationships do exist, the attitudes of teachers are for the most part independent of student achievement. *Expectations* are quite closely tied to student achievement, but *attitudes* appear to be more closely related to the personal qualities of the student and to his reaction to the teacher. Thus, a high achiever is not necessarily going to be liked nor is a low achiever necessarily going to be rejected. Depending upon the student's response to the teacher, a high achiever can just as easily be treated with indifference, and a low achiever can easily become the object of teacher concern rather than teacher rejection.

REFERENCES

Brophy, J., & Good, T. *Teacher-student relationships: causes and consequences.* New York: Holt, Rinehart & Winston, 1974.

Feshbach, N. Student teacher preferences for elementary school pupils varying in personality characteristics. *Journal of Educational Psychology*, 1969, **60**, 126–132.

Good, T., & Brophy, J. Behavioral expression of teacher attitudes. *Journal of Educational Psychology*, 1972, **63**, 617–624.

Jenkins, B. *Teachers' views of particular students and their behavior in the classroom.* Unpublished PhD dissertation, The University of Chicago, 1972.

Silberman, M. Behavioral expression of teachers' attitudes toward elementary school students. *Journal of Educational Psychology*, 1969, **60**, 402–407.

Willis, S. *Formation of teachers' expectations of students' academic performance.* Unpublished PhD dissertation, The University of Texas at Austin, 1972.

Yarrow, M., Waxler, C., & Scott, P. Child effects on adult behavior. *Developmental Psychology*, 1971, **5**, 300–311.

(Received August 27, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 530–533

DIFFERENTIAL EFFECTS OF REWARD VERSUS NO-REWARD INSTRUCTIONS ON THE CREATIVE THINKING OF TWO ECONOMIC LEVELS OF ELEMENTARY SCHOOL CHILDREN[1]

ROGER A. JOHNSON[2]

Old Dominion University

This study examined (*a*) the effects of immediate- and delayed-reward instructions on performance on the Figural Form A of the Torrance Tests of Creative Thinking and (*b*) the possible interactions between reward conditions and economic status and between reward conditions and grade level. Subjects who received immediate- and delayed-reward instructions scored approximately the same and significantly higher than subjects who did not receive any reward instructions. There were no main effects for grade level or economic status. The only significant interaction was reward condition with economic status. The performance of disadvantaged subjects was significantly higher under both reward conditions, while the performance of the relatively advantaged subjects was slightly higher in the no-reward condition.

The present study is focused on the differential effects of immediate- versus delayed-reward instructions on the creative thinking of two economic levels of elementary school children. Ward, Kogan, and Pankove (1972) did not find any difference in the fluency dimension of creative thinking when using immediate and delayed rewards. However, they suggested that there may be a difference in using immediate and delayed rewards when the criteria of creativity are extended to include originality, flexibility, or elaboration. Furthermore, since it would be almost impossible to reward the subject after each original, flexible, and elaborate response, it was more practical to manipulate reward instructions than to reward the subject after each creative response.

Another gap in knowledge lay in the differential effectiveness of reward conditions on economic class with creative thinking as the dependent variable. Although no studies have been done on the relationship between reward, scores on a creativity test, and economic status, several studies have been published on delayed gratification as it relates to different socioeconomic classes.

Investigators are divided on whether middle-class subjects make more delayed-reward selections than lower-class subjects. Studies by Levy (1968), Maitland (1966), and Walls and Smith (1970) indicated that middle-class subjects make more delayed-reward selections than lower-class subjects. On the other hand, Shipe and Lazare (1969) failed to find any significant differences in the ability of different socioeconomic classes to delay gratification.

The present study explored whether there were differences in the way members of different economic classes responded to rewards, with creative thinking as the dependent variable. The questions, Whether reward conditions interact with the child's grade level and whether any differences exist

[1] This article is based on a doctoral dissertation submitted to the University of Georgia in 1973. The author is grateful to Paul Torrance, his major professor, for his guidance and assistance.

This article has been accepted for Southeastern Psychological Association, 1974.

[2] Requests for reprints should be sent to Roger A. Johnson, School of Education, Old Dominion University, Norfolk, Virginia 23508.

on the Figural Form A of the Torrance Tests of Creative Thinking (TTCT; Torrance, 1966) between economic classes and/or grade levels, were also investigated.

Method

Design and Statistics

The three-group, randomized subjects, posttest-only design was used (Campbell & Stanley, 1966). Subjects in Grades 3, 4, and 5 were assigned to one of two experimental groups or to a control group by reference to a table of random numbers.

The immediate-reward group was told that if they worked hard on the game, they would each receive six prizes following the game. The delayed-reward group was given the same instructional set except that they were told they would have to wait one week to receive the prizes. After the students understood the instructions, the test was administered according to the standard instructional set as found in the *Directions Manual and Scoring Guide* for the *TTCT* Booklet A (Torrance, 1966). The control group was administered the test under identical conditions with no mention of reward.

Scores on Fluency, Flexibility, Originality, and Elaboration of the *TTCT* were the dependent variables. The Fluency score represents the total number of relevant responses produced; Flexibility, the number of different approaches or categories found in the responses; Originality, the unusualness and creative strength of the responses; and Elaboration, the number of different ideas produced to elaborate responses. The reliability and validity of the *TTCT* is extensive and satisfactory (Torrance, 1966, 1972). The data were analyzed using unweighted means analysis and a $3 \times 3 \times 2$ analysis of variance design (Glass & Stanley, 1970). Posttest scores were subjected to an analysis of variance to estimate main and interaction effects. The predetermined level of significance was .05.

Subjects and Data Collection

One-hundred and forty-five children (84 whites and 61 blacks) who attended Grades 3, 4, and 5 at Comer Elementary School in rural Georgia served as subjects. Fifty-two subjects whose family income qualified them for a free lunch were classified as disadvantaged. Ninety-three subjects not qualified for a free lunch were classified as relatively advantaged. Slightly more blacks than whites were classified as disadvantaged.

Subjects within each grade level were divided into two groups, namely, disadvantaged and relatively advantaged, with each subgroup further divided at random into three treatment levels (immediate reward, delayed reward, no reward).

Because of the test author's recommendation that disadvantaged children react better when tested in small groups and because it would be easier to assist subjects in small groups to record picture titles, the researcher tested each of the two economic subgroups within the treatment levels in turn. Six white, trained graduate student assistants were assigned at random to administer the test to each of the six tested groups to control for the extraneous tester variable.

Results and Discussion

The means and standard deviations for Fluency, Flexibility, Originality, and Elaboration are shown in Table 1. The fifth-grade subjects had the highest means on all four dimensions of the TTCT. The fourth- and fifth-grade subjects had approximately the same means for Flexibility and Originality; the third and fourth graders had essentially the same scores on the Fluency dimension. Elaboration scores increased steadily from the third through the fifth

TABLE 1

Means and Standard Deviations for Main Effect Subsets: Fluency, Flexibility, Originality, and Elaboration

Factor/level	*n*	Fluency		Flexibility		Originality		Elaboration	
		$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
Grade level									
Third grade	56	20.43	4.15	13.42	2.83	20.92	4.75	34.60	8.06
Fourth grade	41	20.91	4.70	15.44	3.25	25.39	7.35	39.02	14.60
Fifth grade	48	22.65	4.55	15.94	2.23	25.71	6.42	41.04	5.75
Reward condition									
Immediate reward	51	23.06	3.99	15.55	2.30	25.22	3.80	43.87	6.00
Delayed reward	44	23.10	3.83	16.76	2.02	28.27	4.34	38.69	12.28
No reward	50	17.83	3.28	12.49	2.63	18.54	6.52	32.10	8.18
Economic status									
Disadvantaged	52	22.08	5.74	14.65	3.59	22.51	7.95	37.13	11.95
Relatively advantaged	93	20.58	2.31	15.21	2.10	25.51	4.02	39.31	8.08

grade. Subjects in the no-reward condition had the lowest scores on all four dimensions of the TTCT. There was no consistent difference between means for the immediate- and delayed-reward groups. The disadvantaged subjects performed slightly better on Fluency, while the advantaged subjects had a slight edge on Flexibility, Originality, and Elaboration.

All four dimensions of creative thinking were subjected to an analysis of variance to determine if the observed differences in means were significant and to examine possible interaction effects. On all four dimensions of Figural Form A of the TTCT, there was a main effect for reward condition, indicating that one or more of the reward conditions were different from each other (Fluency, $F = 7.89$, $df = 2/127$, $p < .01$; Flexibility, $F = 7.98$, $df = 2/127$, $p < .01$; Originality, $F = 8.08$, $df = 2/127$, $p < .01$; Elaboration, $F = 3.88$, $df = 2/127$, $p < .05$). A Newman-Keuls multiple-comparison test indicates that the immediate-and delayed-reward conditions were not significantly different from each other. Subjects in the no-reward condition scored significantly lower than subjects in the other two reward conditions. There were no significant differences in the main effects for grade level and economic status.

The only significant interaction was reward condition with economic status, which is depicted for Fluency in Figure 1. The performance of disadvantaged subjects was significantly higher under both reward conditions, while the performance of the relatively advantaged subjects was slightly higher in the no-reward condition. The interactions for Flexibility, Originality, and Elaboration were also significant and quite similar to the graph depicted in Figure 1 (Fluency, $F = 7.43$, $df = 2/127$, $p < .01$; Flexibility, $F = 3.68$, $df = 2/127$, $p < .05$; Originality, $F = 3.92$, $df = 2/127$, $p < .05$; Elaboration, $F = 4.39$, $df = 2/127$, $p < .05$).

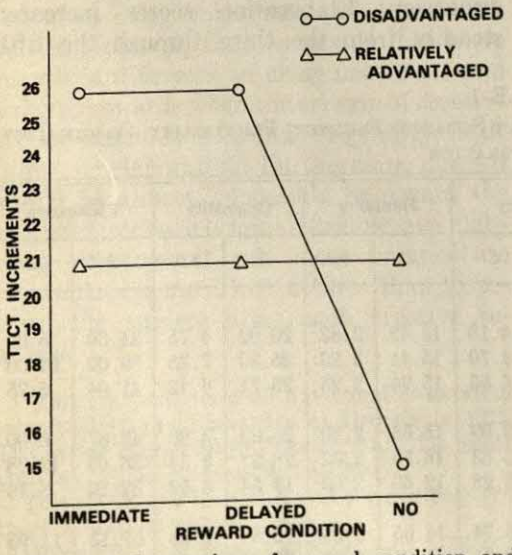


FIGURE 1. Interaction of reward condition and economic status for Fluency.

The intercorrelations of the four TTCT scores were as follows: Fluency with Flexibility was .72; Fluency with Originality was .65; Fluency with Elaboration was .25; Originality with Elaboration was .25; Originality with Flexibility was .31; and Flexibility with Elaboration was .38. All intercorrelations of the four TTCT scores reached significance ($df = 143$, $p < .01$).

It is an accepted postulate of social learning theory that behavior potential is a function of the reinforcements that the subject expects to receive as a result of engaging in that behavior. In the present study, receiving prizes for "working hard" on a creativity test appealed mostly to the disadvantaged subject. Apparently, the consequences of working hard (prizes) were viewed as being worth the effort only by the disadvantaged. This might be attributable to less access to such prizes and, hence, less satiation.

The fact that reward instructions were differentially effective with different economic classes has an important educational implication. If the economic status of the subject is an integral part of the motivational techniques that can be used successfully, then it is important to determine exactly which reinforcers are most effective with each economic class. Once this has been determined, school programs can be developed in accordance with the unique reinforcement patterns of the subjects enrolled.

REFERENCES

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1966.

Glass, G. V., & Stanley, J. C. *Statistical methods*

in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Levy, M. F. *Tolerance for deferred gratification and social class: An experimental investigation of some alternative explanations of differential preference for immediate reward*. (Doctoral dissertation, New York University) Ann Arbor, Mich.: University Microfilms, 1968, No. 69-11, 820.

Maitland, S. D. *Time perspective, frustration-failure and delay of gratification in middle-class and lower-class children from organized families*. (Doctoral dissertation, University of Minnesota) Ann Arbor, Mich.: University Microfilms, 1966, No. 67-866.

Shipe, D., & Lazare, S. Two measures of delay of gratification: Age and socioeconomic status in young children. *Proceedings of the 77th Annual Convention of the American Psychological Association*, 1969, **4**, 461–462. (Summary)

Torrance, E. P. *The Torrance tests of Creative Thinking: Norms-technical manual*. Lexington, Mass.: Personnel Press, 1966.

Torrance, E. P. Predictive validity of the Torrance Tests of Creative Thinking. *Journal of Creative Behavior*, 1972, **6**, 236–252.

Torrance, E. P., & Myers, R. E. *Creative learning and teaching*. New York: Dodd, Mead, 1970.

Walls, R. T., & Smith, T. S. Development of preference for delayed reinforcement in disadvantaged children. *Journal of Educational Psychology*, 1970, **61**, 118–123.

Ward, W. C., Kogan, N., & Pankove, E. Incentive effects in children's creativity. *Child Development*, 1972, **43**, 669–676.

(Received August 28, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 534–539

EFFECTS OF THE AVAILABILITY OF OBJECTIVES AND/OR RULES ON THE LEARNING PROCESS[1]

PAUL F. MERRILL[2]

Florida State University

The effects of behavioral objectives and/or rules on the learning process were investigated using a computer-assisted learning task. The 130 college subjects were administered six ability tests and randomly assigned to an example-only, an objective–example, a rule–example or an objective–rule–example treatment. The subjects were required to meet a minimum criterion performance at each level of the task. Rules significantly reduced the number of examples and the total time required to complete the task and increased performance on a transfer test. Objectives did not significantly affect total or display latency but significantly reduced test-item response latency and the required number of examples. Objectives and/or rules also significantly reduced the requirement for reasoning ability.

It seems that even though educational psychologists (Bloom, 1956; Bobbit, 1924; Tyler, 1951) had been stressing the need for precise statements of instructional objectives for many years, it was not until Mager (1961) published his book on preparing objectives that the educational community started to take instructional objectives seriously. Since Mager's book, many people have mounted the bandwagon and filled the literature with articles extolling the virtues of instructional objectives. However, there are those (Ebel, 1967; Eisner, 1967a; Jackson & Belford, 1965; Kliebard, 1968) who question the value of objectives and feel they might actually be a hindrance to the design of instruction. After an interchange of views in the literature, Eisner (1967b) responded to his critics by pointing out that the contribution of educational objectives to curriculum construction, teaching, and learning is an empirical problem, while most articles that have been written are merely logical arguments. He further claims that the little research that has been done is at best inconclusive. In a recent review of current research on the effects of presenting behavioral objectives to students, Duchastel and Merrill (1973) suggest that future research should focus on interactions between the availability of objectives and both task characteristics and individual differences.

The purpose of this study was to investigate what effects the presentation of behavioral objectives would have on the learning process. Specifically, this study was conducted to further clarify (*a*) how the presentation of objectives would affect subjects' performance on criteria measures; (*b*) how other task characteristics would interact with the presentation of objectives; and (*c*) how individual aptitudes would interact with the presentation of objectives.

It was hypothesized that objectives would serve as orienting stimuli which dispose the student to attend to, process, and organize relevant aspects of displayed information in accordance with the stated objectives. Therefore, the presentation of objectives was expected to reduce the number of examples and the amount of time required to learn the task, facilitate performance on transfer–retrieval criterion measures, and reduce the requirements for memory and

[1] This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-67-A-0126.

[2] Requests for reprints should be sent to Paul F. Merrill, Center for Educational Design—Computer Support of Instruction, Room 10, Kellum Hall, Florida State University, Tallahassee, Florida 32306.

reasoning abilities. However, it was expected that these effects would be tempered by or interact with other properties of the learning task. If objectives are inserted in a task that has minimal orienting or organizing stimuli, then the above hypothesized effects should be very evident. On the other hand, if objectives are inserted in a task which has other effective orienting stimuli such as rules, then the objectives would be somewhat redundant and have a more subtle effect.

Method

Subjects

The 160 subjects who participated in this study were taken from four sections of an introductory educational psychology course and three sections of a science education course at the University of Texas at Austin. All subjects were required to participate as a class assignment. However, 30 of the original subjects were eliminated because they failed to complete all three phases of the study.

Ability Measures

A battery of six cognitive ability tests was administered to all subjects. The battery consisted of three tests selected from the *Kit of Reference Tests for Cognitive Factors* (French, Ekstrom, & Price, 1963) and three task-relevant tests developed for this study. The task-relevant tests required the subjects to process the same type of information that must be processed in the learning task, while the published tests required similar processes on information not related to the task. A list of the individual tests and their factor designations appears in Table 4.

Experimental Tasks and Materials

The learning task used in this study consisted of a hierarchical imaginary science called the Science of Xenograde Systems. The structure and content of the task were similar to those of formal science topics, but the imaginary nature of the science assured that none of the subjects had any previous experience with the task. In the version of the science used for this study, a Xenograde system consists of a nucleus with an orbiting satellite. The satellite is composed of small particles called alphons which also may reside in the nucleus. The subject matter of the science deals with the principles or rules by which the activity of satellite and alphons may be predicted. The terminal objective of the task requires that subjects predict and record the state of the alphons and satellite of a Xenograde system at successive time intervals given the initial state of the system at time zero.

The instructional program consisted of 10 modules. The materials for each module included (*a*) a statement of a subobjective, (*b*) a statement of a rule, (*c*) five examples of the rule, and (*d*) five short, constructed response tests. The examples were in the form of partial Xenograde tables which showed the activity and relationships of a Xenograde system at several points in time.

The instructional program was written in the Coursewriter II language and presented to the subjects by the IBM 1500/1800 computer-assisted instruction system.

Procedure

After the administration of the cognitive ability test battery, the subjects were randomly assigned to one of four groups: an example-only group ($n = 32$), an objective–example group ($n = 33$), a rule–example group ($n = 32$), or an objective–rule–example group ($n = 33$).

In learning the science, subjects in the example-only group received an example of the first rule of the science displayed on a cathode ray tube. After studying the example, each subject responded to a three-item, constructed response test which required him to predict certain values using the rule inferred from the example. If the subject responded correctly to two out of the three test items, he was given an example of the next rule in the sequence. Otherwise, he was given another example of the same rule followed by another three-item test. This sequence of new examples followed by a test continued until the subject responded correctly to two of the three test items or received five examples. The task was completed after all 10 rules of the science were learned to the required criterion. A posttest was administered immediately following completion of the learning task, and retention and transfer tests were administered two weeks later.

The subjects in the other three groups learned the science by the same basic procedure except for the following treatment differences. The objective–example group was shown a statement of a subobjective on an image projector while the corresponding example was displayed on a cathode ray tube. The rule–example group was shown a statement of the rule corresponding to each example, and the objective–rule–example group received both the objective and the rule in addition to the example.

Results

In addition to total scores on the six cognitive ability tests, posttest, retention test, and transfer test mentioned in the previous section, data were obtained for each subject on the following criteria: total number of examples required to learn the science, display latency, test-item response latency, and total latency. Display latency was the total time the subject spent studying the examples

TABLE 1

GROUP MEANS AND STANDARD DEVIATIONS OF THE NUMBER OF EXAMPLES REQUIRED TO LEARN THE TASK

Objectives	Rules			
	No		Yes	
	M	*SD*	*M*	*SD*
No	15.0	3.6	11.0	2.0
Yes	13.3	3.4	10.6	1.2

and, depending upon the subject's treatment group, the corresponding rules and/or objectives. Test-item response latency was the total time required by the subject to respond to the three-item tests following each example display. Total latency was merely the sum of the display and the test-item response latencies.

The descriptive statistics and reliabilities of the ability tests, posttest, retention test, and transfer test are reported elsewhere (Merrill, 1970).

The group means and standard deviations for the number of examples required to learn the task are presented in Table 1. The results from a two-factor analysis of variance revealed a significant rule effect ($F = 48.7$, $df = 1/126$, $p < .001$) wherein the presentation of rules reduced the number of examples required to learn the task. A significant objective effect ($F = 4.7$, $df = 1/126$, $p < .05$) shows that the presentation of objectives also reduced the number of examples required, but this reduction was not nearly as marked as the reduction caused by the presentation of the rules. An examination of the group frequency distributions with number of examples as criterion revealed that 73% of the subjects in the objective–rule–example group, 62% of the subjects in the rule–example group, and 21% of the subjects in the objective–example group learned the task in a minimum number (10) of trials. In contrast, only 3% of the subjects in the example-only group learned the task in 10 trials.

The means and standard deviations for each group on the three latency measures may be found in Table 2. The latency measures were also analyzed using a two-factor analysis of variance. A significant rule effect was obtained on all three measures ($F = 21.9$, $df = 1/123$, $p < .001$, for display latency; $F = 48.8$, $df = 1/126$, $p < .001$, for test-item response latency; and $F = 39.2$, $df = 1/123$, $p < .001$, for total latency), with the rule groups taking considerably less time to study the displays and to respond to the criterion items. The objective effect was significant ($F = 12.8$, $df = 1/126$, $p < .001$) only on test-item response latency, with the objective groups requiring less time to respond to the test items than nonobjective groups. There also was a significant interaction ($F = 4.2$, $df = 1/126$, $p < .05$), with test-item response latency as criterion. This interaction indicates that the objectives had a greater effect in reducing response latency when added to a task which had no other focusing or organizing stimuli than they did when added to a task which had other effective orienting stimuli such as rules. In other words, the difference in response latency between the example–only and objective–example groups was greater than the corresponding difference between the rule–example and objective–rule–example groups.

Since the experimental procedure required all subjects to perform at a minimum criterion level on each rule before proceeding to the next rule, no group mean differences were expected on the posttest. The confirmation

TABLE 2

GROUP MEANS AND STANDARD DEVIATIONS FOR DISPLAY LATENCY, TEST-ITEM RESPONSE LATENCY, AND TOTAL LATENCY IN SECONDS

Group	Latency					
	Display		Test-item response		Total	
	M	*SD*	*M*	*SD*	*M*	*SD*
Example–only	821.2	379.8	923.3	430.7	1772.8	771.7
Objective–example	865.8	377.5	649.3	253.9	1513.5	569.5
Rule–example	543.9	208.6	493.7	211.7	1037.6	399.1
Objective–rule–example	634.3	218.2	419.8	128.2	1053.5	321.2

TABLE 3

GROUP MEANS AND STANDARD DEVIATIONS FOR POSTTEST, RETENTION TEST, AND TRANSFER TEST SCORES

Group	Posttest		Retention test		Transfer test	
	M	*SD*	*M*	*SD*	*M*	*SD*
Example-only	45.5	4.9	44.2	7.2	11.0	2.8
Objective–example	44.3	11.7	43.3	13.9	13.2	4.4
Rule–example	45.1	12.9	43.6	14.2	14.1	5.9
Objective–rule–example	47.8	12.2	46.2	12.2	14.7	5.0

of this expectation made it possible to attribute any group differences on retention or transfer to the differential treatments rather than to differential posttest performance. The posttest, retention test, and transfer test means and standard deviations may be found in Table 3. Even though the rule groups received significantly fewer examples and took significantly less time to learn the task, their performance on the transfer test was significantly higher than that of the no-rule groups ($F = 7.8$, $df = 1/126$, $p < .01$). The objective effect did not reach significance at an acceptable level, but it did approach significance ($F = 3.1$, $df = 1/126$, $p < .10$), with the objective groups obtaining higher mean transfer scores than the no-objective groups. However, there were no significant group mean differences on the retention test.

The battery of cognitive ability tests was factor analyzed, but consistent with previous findings (Bunderson, Olivier, & Merrill, 1971), it was not possible to separate the factors of induction and general reasoning. Therefore, a two-factor varimax solution which yielded the factors of reasoning and associative memory is presented in Table 4. The reasoning factor is marked by the two induction and the two general reasoning tests.

Regression analyses of the individual ability tests scores, factor scores, and the criterion measures were conducted. A significant Ability × Treatment interaction ($F = 3.16$, $df = 3/122$, $p < .05$) was obtained using test-item response latency as the criterion measure and reasoning factor scores as the covariable. Test-item response latency had a high negative relationship to reasoning, as defined by the reasoning factor scores, for subjects in the example-only group. However, the corresponding relationship between reasoning factor scores and test-item response latency was significantly reduced ($F = 9.28$, $df = 1/124$, $p < .01$) for subjects in the other three treatments. Similar ability by treatment interactions were obtained using individual reasoning test scores as covariables with test-item response and total latency as criterion.

DISCUSSION

The design of the present study was such that all subjects were required to reach a minimum criterion performance at each level of the task before they were allowed to go on to the next level. This procedure was used to assure that all treatment groups would perform at the same level on the posttest. Unless all groups learned the original task equally well, differential performance on retention or transfer measure could not be interpreted in terms of the organization or structure provided by the instructional treatments. The results confirmed the expectation of nonsignificant group differences on posttest performance.

Since there was a negligible decrement in performance between the posttest and retention tests for all treatment groups, the retention interval of two weeks may have been too short for the treatments to have had an effect on retention. However, contrary to the learning by discovery hypothesis (Bruner, 1961), the presentation of rules

TABLE 4

VARIMAX ROTATION FACTOR MATRIX

Tests	Factor loadings	
	Reasoning factor	Associative memory factor
Memory of Number Series	.1877	.8336
First and Last Names Test	.0078	.8465
Bi-Column Number Series	.6001	.0802
Letter Sets	.7006	.1954
Tote Mobile	.7458	.1607
Ship Destination	.8191	−.1103

facilitated performance on the transfer task. Even though the rule groups received significantly fewer examples and took significantly less time to learn the task, their performance on the transfer test was significantly higher than that of the no-rule groups. It seems that precisely stated rules have a greater effect on transfer retrieval than objectives. The weak objective effect may have been due to the fact that the objectives-only specified that transfer retrieval would be required to solve new problems using previously demonstrated relationships.

An examination of the group frequency distributions with number of examples as criterion showed that the presentation of rules enables most subjects to learn the science in a minimum number (10) of trials and, therefore, with nearly zero errors. Objectives had a similar but less pronounced effect. Since the rule treatments brought such a high percentage of subjects to perfect performance in terms of the number of examples required, the full impact of these treatments, using number of examples as criterion, was indeterminate. However, the within-group variance was not similarly restricted in the latency criterion measures.

The hypothesis that the presentation of rules would reduce the amount of time required to learn the task was supported by significant rule effects on all three latency measures. The presentation of objectives did not have the hypothesized effect of reducing the total time required to complete the task. This result would seem to contradict the argument that objectives have a focusing effect if it were not for the reduction in the number of examples required by the objective treatments. A comparison of the component latency measures, display latency and test-item response latency, revealed that objectives either increased or had no effect on display latency but significantly reduced test-item response latency. Apparently, the presentation of objectives affected the efficiency and effectiveness of the subject's information processing and thereby facilitated his performance on the criterion test items.

The presentation of objectives and/or rules did significantly reduce the relationship between reasoning factor scores and test-item response latency. Why the treatments interacted with reasoning abilities using test-item response latency and total latency as criteria and did not interact significantly with display latency and number of examples as criteria is not clear. Apparently, reasoning abilities are more crucial during those stages of the task where subjects respond to the criterion test items.

The hypothesis that objective effects would be greater between the example-only and objective groups than between the rule and rule–objective groups was only supported by the significant interaction found with test-item response latency as criterion. However, an examination of the means for the other criteria shows that the corresponding differences between the means are consistent with the hypothesis. Thus, it is impossible to make broad or general statements about the effect of objectives on the learning process without taking into account the other stimulus properties of the task.

On the basis of the results of this study, it was concluded that objectives have orienting and organizing effects which dispose students to attend to and organize relevant information and thus facilitate performance on criterion-test items constructed in accordance with the objectives. However, these effects are not as pronounced when the learning task contains other orienting stimuli such as rules.

REFERENCES

Bloom, B. S. (Ed.) *Taxonomy of educational objectives, Handbook I: Cognitive domain.* New York: Longmans Green, 1956.

Bobbitt, F. *How to make a curriculum.* Boston: Houghton Mifflin, 1924.

Bruner, J. S. The act of discovery. *Harvard Educational Review,* 1961, **31,** 21–32.

Bunderson, C. V., Olivier, W. P., & Merrill, P. F. *The interaction of reasoning and memory abilities with rule-example vs. discovery instruction in learning an imaginary science.* (Tech. Rep. 3) Austin: University of Texas, Computer-Assisted Instruction Laboratory, 1971.

Duchastel, P. C. & Merrill, P. F. The effects of behavioral objectives on learning: A review of empirical studies. *Review of Educational Research,* 1973, **43,** 53–69.

Ebel, R. L. Some comments. *School Review,* 1967, **75,** 261–266.

Eisner, E. W. Educational objectives: Help or hindrance. *School Review,* 1967, **75,** 250–260. (a)

Eisner, E. W. A response to my critics. *School Review,* 1967, **75,** 277–282. (b)

French, J. W., Ekstrom, R. B., & Price, L. A. *Manual for kit of reference tests for cognitive factors.* Princeton, N. J.: Educational Testing Service, 1963.

Jackson, P. W., & Belford, E. Educational objectives and the joys of teaching. *School Review,* 1965, **73,** 267–291.

Kliebard, H. M. Curricular objectives and evaluation: A reassessment. *High School Journal,* 1968, **51,** 241–247.

Mager, R. F. *Preparing objectives for programmed instruction.* San Francisco, Calif.: Fearon, 1961.

Merrill, P. F. *Interaction of cognitive abilities with availability of behavioral objectives in learning a hierarchical task by computer-assisted instruction.* (Tech. Rep. 5, Advanced Research Projects Agency Order No. 1261.) Austin: University of Texas, Computer-Assisted Instruction Laboratory, 1970.

Tyler, R. W. *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press, 1951.

(Received September 8, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 540–547

INFLUENCE OF MODE OF PRESENTATION, ETHNICITY, AND SOCIAL CLASS ON TEACHERS' EVALUATIONS OF STUDENTS[1]

MARY JENSEN[2]
Dalhousie University, Nova Scotia

LAWRENCE B. ROSENFELD
University of New Mexico

After seeing, hearing, or seeing and hearing videotapes of middle- and lower-class Anglo, Black, and Chicano students, teachers evaluated students on a set of 15 semantic-differential scales dealing with teachers' classroom evaluative criteria. A $3 \times 3 \times 2$ analysis of variance for repeated measures revealed significant F tests for all 15 scales. Patterns in the data showed that Anglos were rated most favorably. Middle-class Anglos and Blacks were rated more favorably than lower-class Anglos and Blacks, respectively; however, class made no difference in the ratings of Chicanos. More cues were available in the audio mode of presentation than in the visual mode.

The aim of this study was to investigate the possible influence of ethnic and social class stereotyping on teachers' judgments of students as well as the way in which these stereotypes might be transmitted. Rosenthal and Jacobson (1968) touched off a great deal of research on expectancy effects with the publication of their controversial book, *Pygmalion in the Classroom*. In their research, Rosenthal and Jacobson manipulated teachers' expectations by providing teachers with information concerning students' capabilities for academic achievement. However, in many classrooms, initial expectations for students' classroom behavior may result from social stereotypes. If stereotypes do influence teachers' expectations for student achievement and classroom behavior, then it follows that teachers will devalue Black and Chicano students in line with their ethnic stereotypes. A number of studies offer support for this idea (Whitehead & Miller, 1972; Williams, Whitehead, & Miller, 1971; Woodworth & Salzer, 1971). Social class, like ethnicity, may also serve as a basis for stereotyping, and numerous investigations have documented the negative stereotypes which portray lower-class students (Becker, 1952; Miller, 1973; Rosen, 1969; Sexton, 1961; Sewell, Haller, & Strauss, 1969).

Ethnic and social class cues are transmitted both visually and vocally. Secord (1958), Clifford and Walster (1973), and Williams, Whitehead, and Miller (1971) offer evidence supporting the idea that ethnic and social class cues are transmitted visually. Buck (1968), Anisfeld, Bogo, and Lambert (1962), Naremore (1971), and Williams and Naremore (1971) provide evidence that ethnic and social class cues are transmitted paralinguistically, that is, through the nonverbal, vocal properties of speech. The aim of the this study was to investigate the effects of mode of presentation, students' ethnicity, and students' social class upon teachers' judgments of students.

Method

Teachers were placed into groups based upon mode of stimulus presentation (i.e., audio, visual, or audio–visual) and were then exposed to videotapes of students from different ethnic and social class backgrounds. After exposure to the stimulus tapes, teachers rated each student on 15

[1] The project reported herein was performed pursuant to a grant from the National Institute of Education, U.S. Department of Health, Education, and Welfare (NIE Grant NE-G-00-3-0039). However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education. No official endorsement by the National Institute of Education should be inferred.

[2] Requests for reprints should be sent to Mary Jensen, School of Physical Education, Dalhousie University, Halifax, Nova Scotia, Canada.

semantic-differential scales dealing with classroom evaluative criteria.

Subjects

Twelve junior high and 4 senior high schools were randomly selected from a list of schools in a large (population = 300,000), southwestern city public school system. One hundred and sixty-eight teachers from these schools either volunteered or were chosen by their principals to participate in the study since school administrators would not allow a random selection of teacher subjects to be drawn. Experimental conditions were randomly assigned to schools.

In order to have equal *ns* for the analysis, the responses of 12 subjects were randomly discarded from two groups leaving a total of 156 subjects or 52 subjects per mode of presentation group. Forty-nine of the subjects were male; 107 were female. One hundred and thirty-one of the teachers were Anglo, 13 were Chicano, 6 listed "other" as their ethnic membership (excluding Anglo, Black, Chicano, and Indian), and 6 did not respond to the question. The teachers' average age was in the response category of 30–39 years and the average number of years of teaching experience was 9.08. Eighty-five of the teachers held at least a bachelor's degree; 66 held at least a master's degree; 3 had no college degree; and 2 had PhDs.

Measuring Instruments

During May 1972, a mail questionnaire was sent out to a randomly selected group of public school teachers. Thirty-seven per cent of the teachers responded, giving lists of the evaluative criteria they used to judge students in the classroom. These lists contained both social and academic criteria. Since most of the concepts were highly evaluative in nature, the semantic differential was selected as the questionnaire format to be used in this study (Darnell, 1970).

Concepts for the 15 semantic-differential scales were chosen from lists generated by public school teachers. The most frequently occurring concepts were submitted to a separate group of public school teachers who judged them for their relevance to the classroom. The 15 concepts which resulted from this process were as follows: participates in class, has a good attitude, exerts a great deal of effort, attends regularly, performs well on tests, is highly motivated, is cooperative, works well independently, is very intelligent, follows directions, is responsibile, is courteous, is very creative, has a good self-concept, and is neat. For the final instrument, six different forms of these 15 concepts were prepared (one form for each of six students appearing on each videotape). On each form, both the polarity and the order to the scales were randomized to reduce order effects. Finally, the order of the forms themselves was randomized.

Because of the evaluative nature of the concepts generated, it was believed that social desirability might influence teachers' responses to the students. As a check for this possibility, the Crowne-Marlowe (1964) Social Desirability Scale was included as an additional dependent measure in the study.

Stimulus Materials

The videotapes used in the study were copied from those used by Williams in previous studies (Williams, Whitehead, & Miller, 1971). Fifth- and sixth-grade boys representing Anglo, Black, and Chicano ethnic groups and middle and lower social classes were filmed in interview situations in which the boys were asked to discuss their favorite television shows and games. Each of the ethnic by social class conditions were represented on each of the three tapes used in the study. Thus, 18 different boys (6 on each tape) served as stimuli. Each boy was individually interviewed by an Anglo female in her mid-twenties and the tapes contained edited portions of these interviews which were approximately two minutes in duration. All of the boys were neatly dressed, most of them in slacks and sport shirts. All of the boys were selected from schools in or near Austin, Texas; therefore, their speech reflected regional variations typical of that area.

Results

A 3 × 3 × 2 design (Mode of Presentation × Ethnicity × Social Class) was analyzed for each separate semantic differential using an analysis of variance for repeated measures (Games, 1972). A probability level of .01 was selected as the basis for determining significant differences in the analysis of variance and in the subsequent Newman-Keuls multiple-comparison procedure which was used as the follow-up to pinpoint the specific cells involved in the effects found. Table 1 gives the results of the analyses of variance.

A visual inspection of the graphed cell means revealed consistencies in the data for each of the interactions obtained. The discussion which follows is based upon patterns which were consistent across scales for each of the significant effects obtained. For a complete analysis of the findings, including the *F* tables and the results of the Newman-Keuls procedures, see Jensen (1973).

Main Effects

Two main effects for ethnicity were obtained (Figure 1). The graphed data indicated that Anglo students were evaluated

TABLE 1
RESULTS

Triple interaction	Double interaction	Main effect
Mode × Ethnicity × Class Scale 1: Participation Scale 2: Attitude Scale 5: Test Performance Scale 6: Motivation Scale 9: Intelligence Scale 14: Self-Concept Scale 15: Neatness	Mode × Ethnicity Scale 3: Effort Scale 4: Attendance Scale 13: Creativity Mode × Class Scale 8: Works Independently Ethnicity × Class Scale 4: Attendance Scale 8: Works Independently Scale 11: Responsibility Scale 13: Creativity	Ethnicity Scale 7: Cooperation Scale 12: Courtesy Class Scale 3: Effort

Note. All results are significant ($p < .01$).

more favorably than Black students and that Black students were evaluated more favorably than Chicanos (note that 1 = most favorable rating; 7 = least favorable rating). Comparison of cell means indicates that for both scales, Anglos were rated significantly higher than Chicanos, and on Scale 7: Cooperation, Anglos were also rated significantly higher than Blacks.

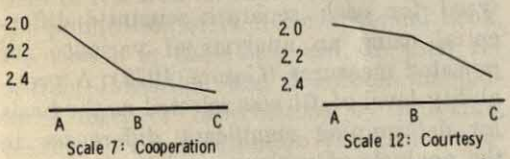


FIGURE 1. Main effects for ethnicity. (Abbreviations: A = Anglo, B = Black, and C = Chicano.)

A main effect for social class was found on Scale 3: Effort (Figure 2). Middle-class students were rated significantly higher than lower-class students on Effort.

Double Interactions

Mode × Ethnicity interactions were obtained on three scales (Figure 3). In regard to mode of presentation, the patterns in the graphed data indicate that Anglos were rated highest in the audio mode and Blacks were rated most favorably in the visual mode; mode of presentation made little difference in the ratings for Chicanos. Mode of presentation seemed to affect ratings for Black students more than ratings for Anglo or Chicano students, with Blacks in the audio mode being rated significantly lower than Anglo students in all three conditions for all three scales. On Scales 4 (Attendance) and 13 (Creativity), Black students in the visual condition were rated significantly higher than Black students in the audio condition. Chicanos, except as already mentioned, were not rated significantly differently than Blacks.

A Mode × Class interaction was obtained on Scale 8: Works Independently. Figure 4 shows that, except in the visual mode, middle-class students were evaluated more favorably than lower class students. However, only the difference in the audio–visual condition was statistically significant.

Ethnicity × Class interactions were obtained for five scales. Figure 5 shows that for Anglo and Black students, teacher evaluations decreased from middle- to

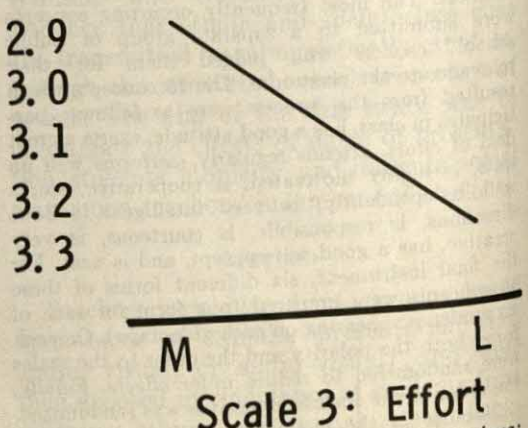


FIGURE 2. Main effect for class. (Abbreviations: M = middle and L = lower.)

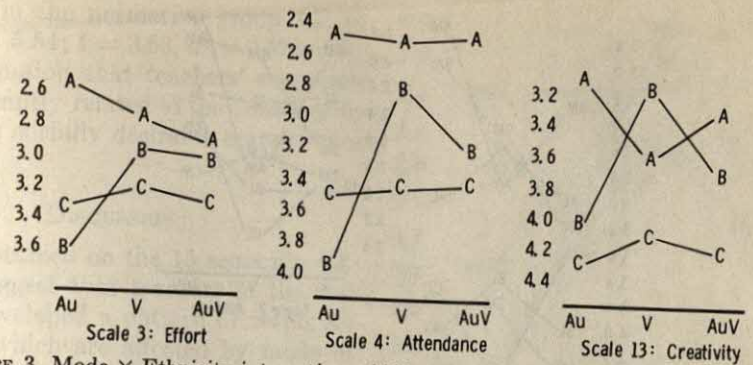


FIGURE 3. Mode × Ethnicity interactions. (Abbreviations: Au = audio, V = visual, AuV = audio–visual, A = Anglo, B = Black, and C = Chicano.)

lower-social-class students. Evaluations for Chicanos did not seem to reflect this trend. Chicano middle- and lower-class students and Black lower-class students were rated

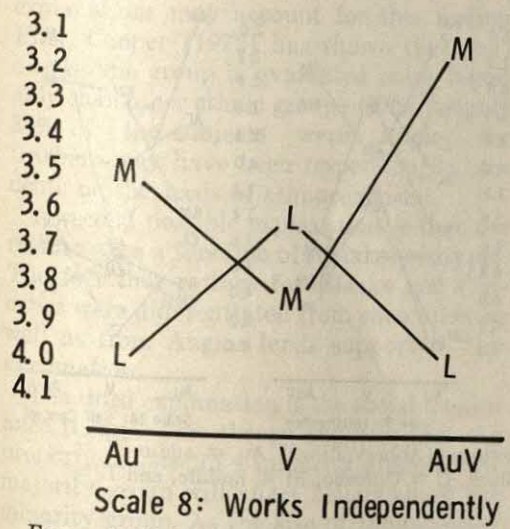


FIGURE 4. Mode × Class interaction. (Abbreviations: Au = audio, V = visual, AuV = audio–visual, M = middle, and L = lower.)

similarly at the lowest end of the scale (i.e., did not differ significantly from each other), while Anglo middle-class students always received the most favorable ratings.

Triple Interactions

Triple interactions were obtained on six scales. The data for all scales except Scale 15 (Neatness) shared patterns. Therefore, what follows is a discussion of the patterns for Scales 1 (Participation), 2 (Attitude), 5 (Test Performance), 6 (Motivation), and 14 (Self-Concept). Scale 15 (Neatness) is discussed separately at the end of this section.

In general, the triple interactions show patterns which are consistent with those obtained with the double interactions (Figure 6). Anglo middle-class students in the audio and audio–visual conditions were rated more favorably than students in nearly every other set of conditions. Anglo lower-class students were rated generally more favorably than both middle- and lower-class Chicanos and lower-class Blacks in all modes (except in the audio–visual mode on Scale 9). Black middle-class students fared better in the visual modes (visual and audio–visual) than in the audio mode, while mode of presentation seemed to make little difference for Black lower-class students. Regardless of mode of presentation and class, Chicanos were represented in the bottom half of the ratings on every scale. With only one exception (visual/Chicano middle-class on Scale 9), none of

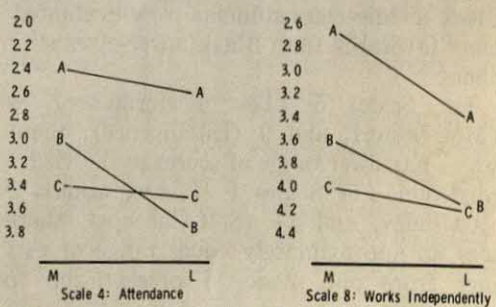


FIGURE 5. Ethnicity × Class interactions. (Abbreviations: A = Anglo, B = Black, C = Chicano, M = middle, and L = lower.)

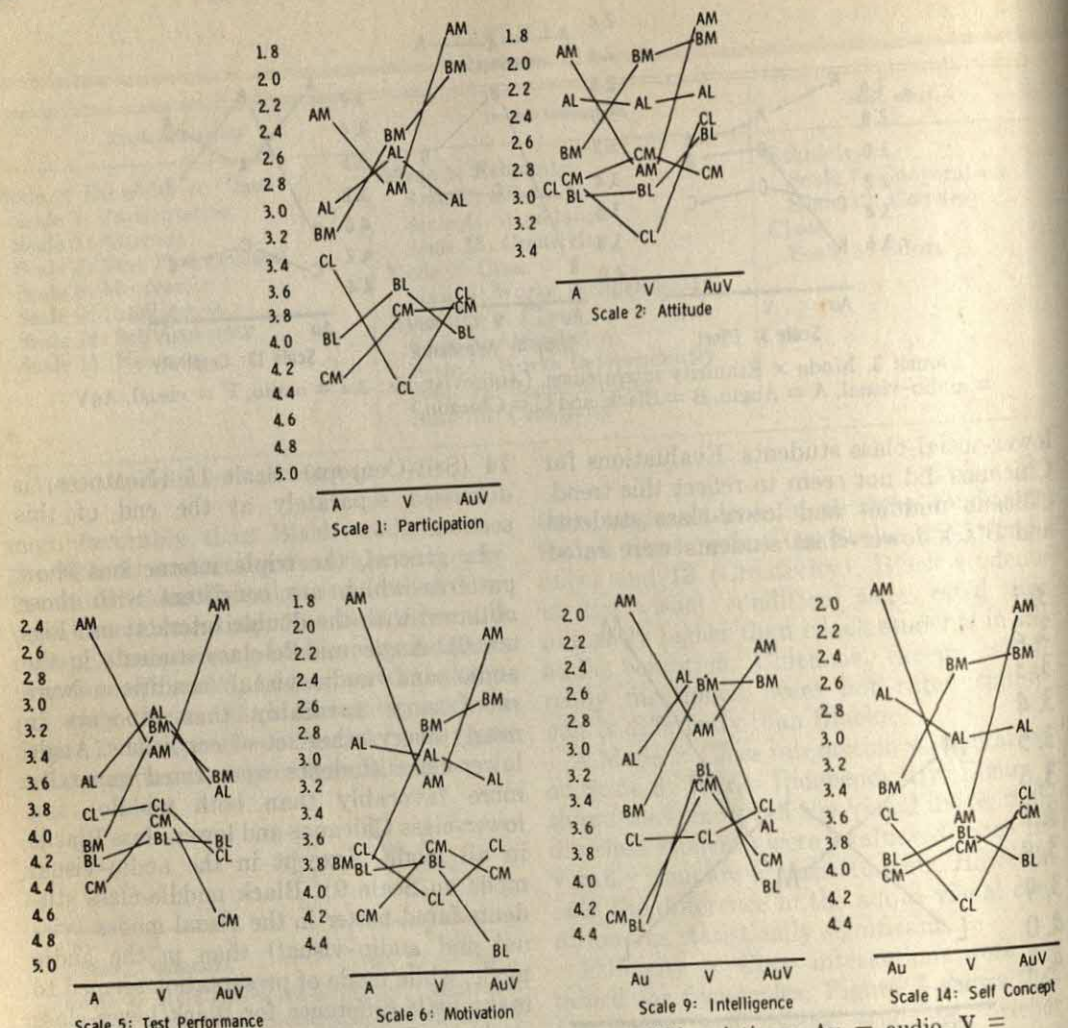


FIGURE 6. Mode × Ethnicity × Class interactions. (Abbreviations: Au = audio, V = visual, AuV = audio–visual, A = Anglo, B = Black, C = Chicano, M = middle, and L = lower.)

the differences for Chicanos were statistically different. Except in the audio mode, Black middle-class students were evaluated more favorably than Black lower-class students.

For Scales 5 (Test Performance), 6 (Motivation), and 9 (Intelligence), there was a narrower range of scores in the visual condition. For Scales 1 (Participation), 2 (Attitude), and 14 (Self-Concept), there was an approximately equal range of ratings from one mode of presentation to another.

Scale 15 (Neatness) did not share patterns with other scales (Figure 7). Mode of presentation had the most obvious effect on this scale since students in the visual condition were rated approximately the same (at the favorable end of the continuum, which could be expected since students were all neatly dressed). Audio cues and audio–visual cues elicited a wider range of ratings than visual cues alone.

Social Desirability

Teachers' responses to the Crowne-Marlow Social Desirability Scale showed that, as a group, teachers in this study had less of a tendency to respond in a socially desirable manner ($\bar{X} = 14.25$, $SD = 5.79$)

than those in the normative group ($\bar{X}$ = 15.99, SD = 5.54; t = 3.63, df = 1573, p < .001). The notion that teachers' responses were significantly related to the tendency to respond in a socially desirable manner was not verified.

Discussion

Results obtained on the 15 semantic differentials suggest that teachers in the test city have developed a pattern of responses to students which are affected by mode of presentation, ethnicity, and social class. The practical result of these stereotypic responses is differential treatment of students. In other words, some students are "more equal" than others.

In general, Anglos were rated higher than either Blacks or Chicanos. Three possible explanations may account for this finding. First, Cooper (1972) has shown that one's own ethnic group is evaluated more favorably than other ethnic groups. Since roughly 84% of the subjects were Anglo, the teachers may have been responding to students on the basis of ethnocentrism.

A second possible explanation is that the ratings were a function of social stereotypes. The fact that ratings for Blacks and Chicanos were differentiated from each other as well as from Anglos lends support to this explanation.

The third explanation is the social dominance theory. This theory suggests that the proportional size of a minority group to the majority group will affect beliefs about the minority group. As the size of the minority group increases, its threat to the social and economic dominance of the majority group increases. In our test city, Chicanos comprise a considerably larger proportion of the population than do Blacks; according to this theory, we would expect Chicanos to be rated lower on the scales than Blacks, and they were. Regardless of the reason(s) operating to produce these results, Anglo students were regarded more favorably than either Black or Chicano students on 13 of the 14 scales which were influenced by ethnicity.

Students' social class, once again proved to be a significant variable in determining

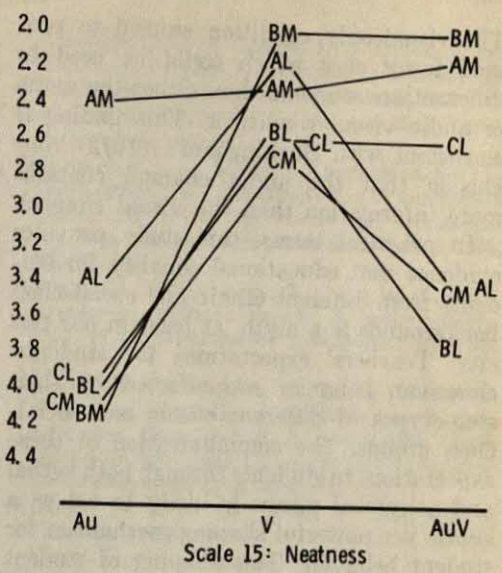


Figure 7. Mode × Ethnicity × Class interaction. (Abbreviations: Au = audio, V = visual, AuV = audio–visual, A = Anglo, B = Black, C = Chicano, M = middle, and L = lower.)

how students were rated. Middle-class Anglo and Black students fared better than lower-class Anglo and Black students. These findings are consistent with stereotypes of the poor as lazy and not caring about getting ahead (Becker, 1952; Davis, 1972) and with results of earlier studies dealing with the evaluation of middle- and lower-class speakers from vocal cues (Harms, 1961; Moe, 1972; Naremore, 1971; Williams, Whitehead, & Traupman, 1971).

However, class made little difference in the ratings for Chicanos, who were consistently rated at the lowest end of the scales. Why ethnicity overrode class is unknown, though the large proportion of Chicanos in the area allows one to speculate that the social dominance theory may have been operating to produce lower ratings for Chicano middle-class students.

Although no main effects for mode of presentation were obtained, mode of presentation did interact with both ethnicity and class to alter the ratings of students. In other words, how students were perceived affected how they were rated. While audio cues helped establish both ethnicity and class, visual cues seemed to be more useful in transmitting ethnic cues than class cues.

The visual-only condition seemed to provide fewer cues which could be used to differentiate students than either the audio or audio–visual conditions. This finding is consistent with Buckingham's (1972) conclusion that the audio channel contains more information than the visual channel.

In practical terms, this study provides evidence that educational equality for students from different ethnic and social-class backgrounds is a myth, at least in our test city. Teachers' expectations for students' classroom behavior are affected by their stereotypes of different ethnic and social-class groups. The communication of these expectations to students through both verbal and nonverbal means is likely to act as a subtle yet powerful shaping mechanism for student behavior. This shaping of student behavior to teacher expectations may account for a large proportion of the differences in educational attainment for different ethnic and social-class groups. An additional finding of this study suggests that Black and Chicano students from lower social classes may be more favorably evaluated if they are seen and not heard. This suggestion comes from the finding that teachers seem to discriminate between students less in the visual mode of presentation than in either the audio or audio–visual modes.

Conclusions

While the following conclusions seem justified in light of the present data, caution should be exercised in generalizing these conclusions to other populations.

1. Anglo students are evaluated more favorably than Black or Chicano students. Chicano students are evaluated least favorably.

2. For Anglo and Black students, class is a salient dimension in teacher evaluations. However, for Chicano students, class has little influence on teachers' ratings, and ethnicity seems to be most relevant to teachers' evaluations.

3. The audio channel contains more information for making classroom judgments than does the visual channel.

4. For Anglo middle-class students, evaluations are more positive when these students are vocal; for Anglo lower-class students, evaluations are more positive in nonvocal situations. Black students are rated more favorably in visual and audio–visual conditions than in the audio-only situation. Regardless of vocal or visual cues, Chicanos are evaluated at the low end of the scale in comparison to Anglo and Black students.

5. Judgments of students are based rarely upon single dimensions. Rather, stereotyping seems to be influenced by a number of considerations, including ethnicity and class. How cues for these two dimensions are received by the teacher (via audio, visual, or audio–visual channels) affects the importance attached to them.

References

Anisfeld, M., Bogo, N., & Lambert, W. I. Evaluational reactions to accented English. *Journal of Abnormal and Social Psychology*, 1962, **65**, 223–231.

Becker, H. S. Social class variations in the teacher-pupil relationship. *Journal of Educational Sociology*, 1952, **25**, 451–465.

Buck, J. The effects of Negro and white dialectal variations upon attitudes of college students. *Speech Monographs*, 1968, **25**, 181–186.

Buckingham, A. T. Language prediction by speakers of English as a first and as a second language: An experimental study. Unpublished doctoral dissertation, Pennsylvania State University, 1972.

Clifford, M. M., & Walster, E. The effect of physical attractiveness on teacher expectations. *Sociology of Education*, 1973, **46**, 248–258.

Cooper, J. G. Perception of self and others as related to ethnic group membership. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

Crowne, D., & Marlowe, D. *The approval motive*. New York: Wiley, 1964.

Darnell, D. K. Semantic differentiation. In P. Emmert & W. D. Brooks (Eds.), *Methods of research in communication*. Boston: Houghton Mifflin, 1970.

Davis, R. D. Situations and language: A sociolinguistic investigation. Paper presented at the Rocky Mountain Foundations of Education Conference, Tucson, December 1972.

Games, P. Analysis of variance for repeated measures. Original program written in Fortran IV, The Pennsylvania State University Computer Center, 1972.

Harms, L. S. Listeners' judgments of status cues in speech. *Quarterly Journal of Speech*, 1961, **47**, 164–168.

Jensen, M. The influence of mode of presentation, ethnicity, and social class on teachers' evaluations of students. Unpublished doctoral dissertation, University of New Mexico, 1973.

Miller, L. M. Evaluational reactions of Mexican-American and Anglo teachers to children's speech. *Western Speech*, 1972, **36**, 109–114.

Moe, J. D. Listener judgments of status cues in speech: A replication and extension. *Speech Monographs*, 1972, **39**, 144–147.

Naremore, R. C. Teacher's judgments of children's speech: A factor analytic study of attitudes. *Speech Monographs*, 1971, **38**, 17–27.

Rosen, B. C. Race, ethnicity, and the achievement syndrome. In B. C. Rosen, H. J. Crockett, Jr., & C. Z. Nunn (Eds.), *Achievement in American society*. Cambridge, Mass.: Schenkman, 1969.

Rosenthal, R., & Jacobson, L. *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston, 1968.

Secord, P. F. Facial features and inference processes in interpersonal perception. In R. Taguiri and L. Petrullo (Eds.), *Person perception and interpersonal behavior*. Stanford, Calif.: Stanford University Press, 1958.

Sewell, W. H., Haller, A. O., & Strauss, M. A. Social status and educational and occupational aspiration. In B. C. Rosen, H. J. Crockett, Jr., & C. Z. Nunn (Eds.), *Achievement in American society*. Cambridge, Mass.: Schenkman, 1969.

Sexton, P. C. *Education and income*. New York: Viking Press, 1961.

Whitehead, J. L., & Miller, L. Correspondence between evaluations of children's speech and speech anticipated upon the basis of stereotype. *Southern Speech Communication Journal*, 1972, **27**, 375–386.

Williams, F., & Naremore, R. C. Social class differences in children's speech: A factor analytic study of attitudes. *Speech Monographs*, 1971, **38**, 17–27.

Williams, F., Whitehead, J. L., & Miller, L. *Attitudinal correlates of children's speech characteristics*. (Final report, Project 0-0336, Grant OEG-0-70-7868(508)) University of Texas, Center for Communication Research, 1971.

Williams, F., Whitehead, J. L., & Traupman, J. Teacher's evaluations of children's speech. *Speech Teacher*, 1971, **20**, 247–254.

Woodworth, W. D., & Salzer, R. T. Black children's speech and teachers' evaluations. *Urban Education*, 1971, **6**, 167–173.

(Received September 12, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 548-553

INITIAL INVESTIGATIONS INTO ACADEMIC SURVIVAL-SKILL TRAINING, DIRECT INSTRUCTION, AND FIRST-GRADE ACHIEVEMENT[1]

HYMAN HOPS[2] AND JOSEPH A. COBB

Center at Oregon for Research in the Behavioral Education of the Handicapped, University of Oregon

Training in academic survival skills was compared with direct instruction via curriculum individualization on reading achievement. Teachers in two classrooms were trained in the contingent reinforcement of group survival skills (GSS); one teacher was given specific instructions in the construction and use of a programmed, individualized reading curriculum (DI); and one teacher was given no new instructions. Children in all classrooms were observed and tested before and after intervention. Both GSS and DI approaches resulted in improved reading achievement when compared with the control group, but only the specific survival-skill training increased the proportion of children's survival-skill behaviors. It was concluded that achievement may be increased directly through curriculum improvements or indirectly through increasing behaviors which are prerequisites for academic achievement.

Continuing investigation into the identification of precise, observable, and manipulable parameters controlling academic achievement is a vital prerequisite for establishing successful classroom intervention procedures. One set of such parameters, academic survival skills, (Cobb, 1970, 1972a) has been found to be correlated with elementary school achievement (Cobb, 1969, 1970, 1972b; Lahaderne, 1968; Myers, Atwell, & Orpet, 1968). These "survival skills" are specific classroom behaviors that are not academic responses per se but are, instead, the necessary basis for such academic responding. Examples include attending to teacher and volunteering to participate in classroom discussions. More recently, empirical evidence has established a clear functional relationship between survival-skill behaviors and reading achievement in the first grade. Increased proportions of survival-skill behaviors resulted in significantly increased scores on standardized achievement tests for children who were initially low on both survival-skill behaviors and achievement (Cobb & Hops, 1973) and for entire regular classrooms (Hops & Cobb, 1973).

Other approaches have also produced increased academic achievement. Children who received direct instruction in reading, particularly with emphasis on the systematic programming of phonics or decoding skills made significant gains in reading (Bliesmer & Yarborough, 1965; Karnes, Hodgins, & Teska, 1968; Waugh, 1969). In an effort to understand the place of survival skills in academic achievement, the present study was designed to determine if the cause-and-effect relationship between achievement and academic survival skills was unidirectional. That is, do increases in achievement effected through a direct attack on the curriculum and academic responding result in concomitant changes in survival-skill levels.

[1] This research was supported by Contract NPECE-70-005, OEC 0-70-4152(607), Bureau of Education for the Handicapped, U.S. Office of Education. Thanks are due to Charles R. Greenwood, Hill M. Walker, and Roberta Taussig for critical readings of the manuscript.

[2] Requests for reprints should be sent to Hyman Hops, CORBEH, 1590 Willamette Street, University of Oregon, Eugene, Oregon 97401.

In evaluating the relationship between academic survival skills, direct instruction, and elementary school achievement, the following hypotheses were tested: (*a*) children who receive direct programmatic instruction in reading increase their achievement but *not* their survival-skill levels; and (*b*) children who receive academic survival-skill training increase their skill levels *both* in achievement and in survival-skill behaviors.

In the present study, the effect of each of the two approaches (survival-skill training and direct programmatic instruction) on first-grade reading achievement was compared with a no-treatment control group.

Method

Subjects

The study was conducted in four first-grade classrooms in a school district of 21,000 pupils. The classrooms were randomly assigned to the experimental conditions. Eighty-one children, for whom complete pre- and postintervention reading achievement scores were obtained, comprised the entire sample. There were 20 pupils in the control (C) classroom, and 19 in the direct instruction (DI) classroom. The two remaining classrooms, both of which received group survival-skill training (GSS), had 21 students each.

Achievement Tests

The Gates MacGinitie Primary A, yielding Comprehension and Vocabulary scores, and the Gates-MacGinitie Reading Readiness tests were administered both prior to and six weeks after the termination of the formal intervention procedures. The mean of the standard scores of the three tests for each student was used as the dependent variable. The number of school days between testings ranged from 72 to 77 days across the four classrooms.

Observations

During the same weeks in which achievement testing occurred, observations were made of the student's classroom behavior during all reading periods for five consecutive days. An interactive coding system (Cobb & Hops, 1971) was used to record the behavior of all the children in a prearranged manner. The observer recorded the behaviors of each student for a minimum of two consecutive eight-second intervals before going on to the next student on the list. After all the children had been coded once, the process was repeated so that the behavior of each child was sampled about the same number of times during each session.

Four full-time and two part-time observers gathered data. Each new observer was required to reach an 85% or higher level of agreement on three occasions with an experienced observer before her independent observations were considered reliable. Following training, the same level of agreement was maintained by periodic reliability checks.

Observer reliability was calculated by dividing the total number of agreements by the total number of agreements and disagreements. Throughout the study, reliability data were collected systematically between observers. Based upon 27 paired observations, reliability ranged from 85% to 99% with a mean of 94%.

Cobb (1970) reported significant correlations between first-grade reading achievement and three academic survival skills: attending (.45), volunteering (.59), and look around (−.41). For the current study, attending was divided into two, newly defined, distinct categories: attending and work. To compute a student's survival-skill level for the pre- and postintervention observation periods, the frequency of look around was subtracted from the summed frequencies of attending, volunteering, and work. The obtained figure was divided by the total frequency of *all* behaviors and represented the proportion of academic survival-skill behaviors in each child's repertoire during each measurement phase.

Intervention

All of the experimental manipulations occurred in the regular classroom setting. The primary focus was alteration of the teacher's behaviors, which were designed to enhance during intervention and maintain during follow-up, favorable academic and/or survival-skill behaviors in the children.

The group survival-skill program (GSS), based upon the work of Packard (1970), Patterson, Cobb, and Ray (1972), Walker, Mattson, and Buckley (1971), and Walker, Fiegenbaum, and Hops (1971), has been described in detail elsewhere (Cobb, 1972a; Cobb & Hops, 1973). Teacher training included the use of modeling, assigned reading, cueing, and daily feedback, all of which were faded out as the teacher acquired the skills and the children's behavior improved. Components of the child-training procedures taught to and employed by the teachers were the pairing of group-nonsocial reinforcement with individual- and group-social reinforcement, vicarious reinforcement, shaping procedures, close monitoring, and the withdrawal of nonsocial reinforcement by gradually increasing the criterion for such reinforcement. The intervention period lasted 20 school days, and the total consultant time for each of the two GSS classrooms was approximately 12 hours.

The approach used in the direct instruction of reading (DI) was based upon the assumption that reading can be taught in a programmatic fashion by individualizing the curriculum. The task (read-

ing) is analyzed into a hierarchy of subskills; the teacher identifies the entering behaviors or particular subskill levels of each child; the child is taught the next task in the hierarchy; and finally, the teacher determines by a preset criterion whether the child has mastered one subskill before proceeding to the next.

Individualizing Reading Instruction (Schmahl, Bowman, & Hops, 1973)[3] was a manual specifically prepared to help the teacher use the procedures. It contained a flowchart which listed in sequential form all the components deemed necessary for the efficient teaching of reading. While it was designed to complement basal curriculums with phonics and vice versa, the greatest emphasis was given to the phonics components. Previous research has shown that beginning reading programs which stress phonics or sound–symbol relationships, even when curriculums are individualized and programmed, are superior to those that do not (Bliesmer & Yarborough, 1965; Bond & Dykstra, 1967; Chall, 1967).

A short written test was administered to the teacher to ensure that she had learned the concepts presented in the manual. Then, using frequent observations, feedback, social praise, and modeling by the consultant, the teacher was first trained to work with three low-reading achievers and was subsequently allowed and encouraged to generalize her training to three similar children and, finally, to the entire classroom. The training period lasted 20 school days, and the consultant spent 17 hours working with the teacher.

In lieu of direct service, the control teacher was provided with a graduate student in special education who acted as a teacher aide during the course of the study. The student was given no specific instructions and spent considerably more time in the control classroom than did any of the consultants in the experimental classrooms.

Results

Preexisting differences between groups on the dependent variable have been discussed as a critical factor in evaluating change scores by Campbell and Stanley (1963). These authors suggest the analysis of covariance as a statistical method of controlling for such differences. Others (O'Conner, 1973) have criticized the use of the analysis-of-covariance model. This criticism is particularly valid when assumptions underlying the model, such as homogeneity of regression coefficients, are not met (Kirk, 1968). In the present experiment, tests for homogeneity of regression were conducted on both dependent variables. The survival-skill data produced an F value of 14.49 ($df = 3/72$, $p < .001$) indicating that the homogeneity assumption was untenable; the achievement data met the requirement with an $F < 1$.

In addition, the initial scores for survival skills and mean achievement (see Table 1) were subjected to one-way analyses of variance to determine if pretreatment differences existed. The F values indicated that there were no significant differences between the groups on the observation ($F = 1.00$, $df = 3/77$, $p > .05$) or achievement ($F = 2.39$, $df = 3/77$, $p > .05$) data.

In the present study, no differences between groups on either variable existed prior to treatment, and the data for one variable could not meet the assumptions underlying the analysis of covariance. Consequently, gain scores for both survival skills and mean achievement were used as the dependent variables and subjected to statistical analyses. Mean scores and standard deviations of both variables are presented in Table 1.

To determine whether the classrooms which received the group survival-skill training made greater gains in survival-skill behaviors than those that did not receive the training, a one-way analysis of variance was carried out on the gain scores. Significant differences between groups were found ($F = 5.00$, $df = 3/77$, $p < .005$). Post hoc comparisons using the Newman-Keuls test indicated that the mean gains for both of the survival-skill-training groups ($GSS_1 = .14$, $GSS_2 = .10$) were significantly greater than the gains made by the direct-instruction classroom (DI = .04) or the control classroom (C = .02). No differences were found between the gain scores of the two GSS classrooms.

A similar analysis was conducted on the achievement gains, that is, on the mean of the three standard scores (Vocabulary, Comprehension, and Readiness). Here, too, a significant F value was obtained ($F = 9.15$, $df = 3/77$, $p < .001$). Post hoc analyses using the Newman-Keuls procedure showed that all three experimental groups

[3] Copies of this manual are available upon request.

TABLE 1

MEAN PRETREATMENT AND MEAN GAIN SCORES FOR SURVIVAL-SKILL BEHAVIORS AND ACHIEVEMENT STANDARD SCORES

Group	*n*	Survival skills				*M* achievement			
		Pretreatment		Gain		Pretreatment		Gain	
		$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
C	20	.63	.09	.02	.14	52.5	7.5	7.6	4.7
GSS_1	21	.59	.05	.14	.08	48.2	6.5	12.9	4.8
GSS_2	21	.55	.14	.10	.12	48.6	5.7	10.9	5.0
DI	19	.59	.14	.04	.13	51.7	5.1	15.0	4.0

Note. Survival-skill scores = attending + work + volunteering − look around; achievement standard scores = (Readiness + Vocabulary + Comprehension) ÷ 3.

[a] Abbreviations: C = control, GSS_1 = Group Survival-Skill Training 1, GSS_2 = Group Survival-Skill Training 2, and DI = direct instruction.

made significantly greater gains than the control classroom (GSS_1 = 12.9, GSS_2 = 10.9, DI = 15.0, C = 7.6). However, the gains made by the curriculum-training classroom were significantly greater than those by GSS_2, but not GSS_1.

The first hypothesis, that children who received direct programmatic instruction in reading would increase their achievement level but not their survival-skill behavior, was supported. The curriculum-training classroom, while showing no gains in survival skills, made significantly greater gains in achievement than both the control group and one of the survival-skill-training groups. The second hypothesis was also supported: children in the group survival-skill-training program demonstrated greater gains in achievement and survival skills than did the children of the control classroom. Thus, the data show that survival-skill behaviors are increased by survival-skill training, and not by training in reading skills. Scores in reading achievement, in contrast, were increased both by direct instruction in reading and also by survival-skill training.

DISCUSSION

The results indicated that increases in reading achievement can be effected *directly* through well-programmed individualized reading instruction and *indirectly* through programs which are designed to increase children's rates of specific classroom behaviors by enabling them to profit more from the existing curriculum. Given an experimentally unmanipulated basal curriculum, as in the survival-skill classrooms, gains in achievement may be established through behavioral improvements in the children. Or, given an experimentally unmanipulated level of survival skills, as in the direct-instruction classroom, achievement increases may be effected through improvement in the curriculum. While not tested in the present study, the most effective program would, presumably, combine both components; children would be taught to attend more and work more (survival skills) to the teacher's presentation of an improved curriculum (direct instruction).

A notable limitation of the present study is the small number of classrooms involved in each of the conditions. Replications of these findings will be required before any firm conclusions can be reached about the relationship between academic survival skills, direct instruction, and academic achievement as measured by standardized achievement tests.

The findings of the present study, while preliminary, have further implications. Both sets of procedures can be of practical importance to the classroom teacher. Teachers with effective curricula may still require help in classroom management for increasing survival-skill behaviors of chil-

dren. On the other hand, children in some classrooms may be working near maximum levels while using a curriculum which severely limits the progress individual children can make. The materials may be too difficult and frustrating for children at the lower end of the achievement distribution or too easy and boring for children at the upper levels.

A second issue bears upon the response cost to the teacher of implementing these procedures. In the authors' experience, programs which require less effort are more likely to be implemented effectively. However, increasing the efficiency of the classroom by using either program can also increase the workload of the teacher as she responds more to individual children's needs. Problems of this nature were beyond the scope of the present study but nevertheless are critical issues.

Another question remains to be answered by additional research. Will curriculum improvements, such as the direct-instruction program used in the present study, have positive benefits only for children above a critical level of survival skills? Presumably, children who do not respond to the teacher's controlling stimuli in a classroom would not respond maximally to the best available curriculum materials. Perhaps, more powerful motivational programs (Hops, Walker, & Hutton, 1973) would be necessary prior to or in conjunction with improved curricula. Further studies are required to determine precisely what the interaction is between a child's survival-skill level, curricula, and increases in academic achievement.

REFERENCES

Bliesmer, E. P., & Yarborough, B. H. A comparison of ten different beginning reading programs in the first grade. *Phi Delta Kappan,* 1965, **56,** 500–504.

Bond, G. L., & Dykstra, R. The cooperative research program in first-grade reading instruction. *Reading Research Quarterly,* 1967, **2**(4), 5–142.

Campbell, D. T., & Stanley, J. S. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963.

Chall, J. *Learning to read: The great debate.* New York: McGraw-Hill, 1967.

Cobb, J. A. The relationship of observable classroom behaviors to achievement of fourth grade pupils. Unpublished doctoral dissertation, University of Oregon, 1969.

Cobb, J. A. *Survival skills and first-grade academic achievement.* (CORBEH Res. Rep. 1) Eugene: University of Oregon, Center at Oregon for Research in the Behavioral Education of the Handicapped, 1970.

Cobb, J. A. Academic survival skills and elementary achievement. In E. Meyen, G. Vergason, & R. Whelan (Eds.), *Strategies for teaching exceptional children.* Denver, Colorado: Love, 1972. (a)

Cobb, J. A. Relationship of discrete classroom behaviors to fourth-grade academic achievement. *Journal of Educational Psychology,* 1972, **63,** 74–80. (b)

Cobb, J. A., & Hops, H. *Coding manual for subject/peer/teacher sequential interactions in academic survival skill settings.* Eugene: University of Oregon, Center at Oregon for Research in the Behavioral Education of the Handicapped, 1971.

Cobb, J. A., & Hops, H. Effects of academic survival skill training on low achieving first graders. *Journal of Educational Research,* 1973, **67,** 108–113.

Hops, H., & Cobb, J. A. Survival behaviors in the educational setting: Their implications for research and intervention. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice.* Champaign, Ill.: Research Press, 1973.

Hops, H., Walker, H. M., & Hutton, S. *CLASS Program for acting-out children.* Eugene: University of Oregon, Center at Oregon for Research in the Behavioral Education of the Handicapped, 1973.

Karnes, M. B., Hodgins, A., & Teska, J. A. An evaluation of two preschool programs for disadvantaged children: A traditional and a highly structured experimental preschool. *Exceptional Children,* 1968, **34,** 667–676.

Kirk, R. E. *Experimental design: Procedures for the behavioral sciences.* Belmont, Calif.: Brooks/Cole, 1968.

Lahaderne, H. M. Attitudinal and intellectual correlates of attention: A study of fourth-grade classrooms. *Journal of Educational Psychology,* 1968, **59,** 320–324.

Meyers, C. E., Atwell, A. A., & Orpet, R. E. Prediction of fifth-grade achievement from kindergarten test and rating data. *Educational and Psychological Measurement,* 1968, **28,** 457–463.

O'Conner, E. F., Jr. Why matching, analysis of covariance, and multiple regression analysis generally fail to control for pre-existing differences. Paper presented at the Annual Meeting of the Western Psychological Association, Anaheim, April 1973.

Packard, R. G. The control of "classroom attention:" A group contingency for complex be-

havior. *Journal of Applied Behavior Analysis*, 1970, **3**, 13–28.

Patterson, G. R., Cobb, J. A., & Ray, R. S. Direct intervention in the classroom: A set of procedures for the aggressive child. In F. W. Clark, D. R. Evans, and L. A. Hamerlynck (Eds.), *Implementing behavioral programs for school and clinics*. Champaign, Ill.: Research Press, 1972.

Schmahl, L., Bowman, J., & Hops, H. *Individualizing reading instruction: A manual for teachers*. Eugene: University of Oregon, Center at Oregon for Research in the Behavioral Education of the Handicapped, 1973.

Walker, H. M., Fiegenbaum, E., & Hops, H. *Components analysis and systematic replication of a treatment model for modifying deviant classroom behavior*. (CORBEH Res. Rep. 3) Eugene: University of Oregon, Center at Oregon for Research in the Behavioral Education of the Handicapped, 1971.

Walker, H. M., Mattson, R. H., & Buckley, N. K. The functional analysis of behavior within an experimental class setting. In W. C. Becker (Ed.), *An empirical basis for change in education*. New York: Science Research Associates, 1971.

Waugh, R. The case for precision in teaching reading. *The Elementary School Journal*, 1969, **70**, 124–129.

(Received September 17, 1973)

Announcement

In addition to the members of the Advisory Board, the *Journal of Educational Psychology* asks other psychologists to aid in the review of manuscripts. In this way we have been able to achieve a broader review base. The Editors wish to thank the following people for their generous assistance to the Journal.

Richard Anderson, *University of Illinois*
Benjamin Bloom, *University of Chicago*
Joel B. Cohen, *National Analysts, Inc., Philadelphia*
E. B. Coleman, *University of Texas*
James Crouse, *University of Delaware*
Shel Feldman, *Swarthmore College*
James Hall, *Northwestern University*
Duncan Hansen, *Florida State University*
C. D. Hardyck, *University of California, Berkeley*
Martin Kaufman, *Department of Education, H. E. W.*
Martin Kling, *Rutgers—The State University*
Paul Kolers, *University of Toronto*
Jacob S. Kounin, *Wayne State University*
Luis M. Laosa, *University of California, Los Angeles*
Richard J. Light, *Harvard University*
Richard H. Lindeman, *Teachers College, Columbia University*
Eleanor E. Maccoby, *Stanford University*
Carol Ann Moore, *Educational Testing Service*
Daniel M. Ogilvie, *Livingston College, Rutgers—The State University*
David A. Potter, *Educational Testing Service*
Thomas J. Quirk, *Principia College*
Marjorie Ragosta, *Educational Testing Service*
Robert R. Rosenthal, *Harvard University*
Julian B. Rotter, *University of Connecticut*
Marshall Smith, *National Institute of Education*
Frederick Strassburger, *American Psychological Association*
Robert J. Thorndike, *Teachers College, Columbia University*
William C. Ward, *Educational Testing Service*
Esther Westervelt, *Simmons College*
Kenneth M. Wilson, *Educational Testing Service*
Allen Yates, *Educational Testing Service*

Journal of Educational Psychology
1974, Vol. 66, No. 4, 554-562

SIGNAL SYSTEMS OF LESSON SETTINGS AND THE TASK-RELATED BEHAVIOR OF PRESCHOOL CHILDREN[1]

JACOB S. KOUNIN[2]
Wayne State University

PAUL V. GUMP
University of Kansas

Thirty-six different teachers were videotaped teaching 596 lessons in a special room. Lessons were seen as signal systems to participants. These signal systems were characterized as varying along the dimensions of continuity, insulation, and intrusiveness. Using task involvement as the criterion, the most successful lessons are those in which there is a continuing and protected signal system (as in individual construction). Lessons of average success are those with a continuous input from a constant source (books, records, teacher demonstrations). The least successful lessons are those dependent upon discontinuous inputs from other children (role play, group construction) and those having high intrusiveness (gross motor activity, loud musical instruments).

A major job of teachers is to create and maintain beneficial activity settings. Yet the amount of information developed by the research community concerning activity settings in schools or in other environments is minute (Moos, 1973). Studies regarding how children develop or what specific interventions can do to their behavior and learning abound, but there is a paucity of research about the impact of activity settings upon child behavior.

Educators specializing in early childhood education appear to have some conventional wisdom about the kinds of activity settings to provide in a preschool or at least some consensus about the kinds of activities to provide. Most preschools, for example, provide large blocks and coloring equipment, and none provide calculus texts or standard-size basketball courts. Shure (1963) has pioneered research showing differences in such variables as popularity, population density, sociality, and constructiveness among different open settings of a preschool. The authors, however, have been unable to locate any research dealing with properties of formal lesson settings in preschools as these relate to the behavior of children participating in the lessons.

This report is part of, and was generated by, a research project investigating the preschool as a complex of activity settings. A school may contain both open settings (block corner, sandbox, art section with easels) as well as prescriptive lesson settings. Each activity setting has two aspects: a particular milieu with its appropriate behavior objects and a format or program with a standing behavior pattern which fits that milieu. Barker (1968) has provided the generic name of *synomorphs* for such ecological units.

The focus of this paper is upon the relationships between properties of formal lessons and the behavior of child participants. Such formal lessons are parts of school environments, and a lesson may be regarded as an activity setting inhabited by teacher and pupils who are *in* the lesson

[1] This investigation is supported by PHS Research Grant MH-15472 from the National Institute of Mental Health. Lawrence Sherman directed the computer programming and provided considerable assistance in a variety of ways. The authors wish to express their gratitude to Norma Law, Sharon Elliott, and Hilda Weems of the Wayne State Nursery School for their intensive help and to the 36 student teachers who participated in this research.

[2] Requests for reprints should be sent to Jacob S. Kounin, Department of Educational and Clinical Psychology, Wayne State University, Detroit, Michigan 48202.

much as they might be in a playground game or in a school play.

The research asks whether certain qualities or dimensions of lessons can be delineated, whether these qualities can predict to the task-related behavior of children in these lessons, and whether these predictions can be made independently of the differences of the teachers and children who inhabit these lessons.

Method

Population, Setting, and Regime

The study took place in a university nursery school in a metropolitan area. One group of 20 children attended a morning session and another group of 20 attended an afternoon session. Each group was heterogeneous with respect to age, sex, race, and socioeconomic background. Two experienced, certified teachers were present at all times. In addition, six student teachers were present, three in the morning and three in the afternoon.

The children, one half boys and one half girls, ranged in age from 29 to 65 months at the time they were first videotaped. The racial mixture consisted of 60% non-Caucasian (including an Indian, an Oriental, and a Mexican) and 40% Caucasian. The incomes of the children's families ranged from below $3,000 a year to $75,000 a year.

As part of the daily routine of the preschool, a preformed subgroup of children from each session was taken to a special room and was presented with a formal lesson by a student teacher. The lesson groups were formed by the head teacher on the basis of approximate chronological age but were heterogeneous with respect to sex and race. The number of children in a lesson group occasionally varied due to absences and ranged from three to nine; however, 87% of the 596 lessons contained from four to eight children.

The lessons occurred in a limited and protected segment of the total nursery school regime. Their milieu was a 2.13 × 2.74 meter room which had one entrance door, one window, and a one-way observation window. A high shelf on the wall, accessible only to the teacher, and a light bulb hanging from the middle of the ceiling were the only permanent furnishings in the room. (The light was switched off by the teacher when the short rest period prior to the lesson was initiated and was switched on when the teacher stopped the rest period and started the lesson proper.) The room was empty except for those supplies that were brought in by the teacher for a specific lesson. A group was scheduled for a lesson period for approximately 20 minutes a day, including the brief rest period.

The lessons were chosen and planned by the student teachers. There was no particular unifying theme from lesson to lesson except, perhaps, judged appropriateness for the preschool-aged child. There was no standard to stress language, concepts, numbers, or any kind of "subject." Rather, the lessons were diverse in content and format both between teachers and for any one teacher on different days.

The lessons were recorded on videotape in order to have enduring and complete information that could be viewed as many times as required. A stationary video camera was attached to a corner about two meters from the floor, and a microphone was hung on the side having the one-way window. A wide-angle lens made it possible to view most of the floor space. The operator and the recording equipment were in a different room, behind the one-way window.

The operator videotaped 596 lessons taught by 36 different teachers (6 teachers per academic quarter; 3 for the morning and 3 for the afternoon groups). All teachers of any academic quarter taught all three groups an equal number of times, and all the lessons of a teacher's first two groups were recorded. The recording extended over a period of two years, and 87 different children were the unrehearsed participants.

The videotaping of the lessons was an established routine for the two years the study was in progress. The investigator had no influence on the lessons taught nor did he provide any feedback to the teachers about their procedures or lessons.

At this point we would like to interject some comments about the general procedure. We started with a curiosity but without any theory except the general notion that settings and lesson formats influence behavior. The approach was ecological in that there were no external constraints either on input data or on output data. Nevertheless, the situation at the nursery school provided an investigator with an opportunity to study uncontrived phenomena that met most of the requirements of experimentally created events. Variations in independent variables occurred with sufficient frequency and magnitude; background variables such as the population and general school situation were constant; the populations participating in the lessons were balanced in that the different teachers met the same groups the same number of times (for any one year, the nine teachers of the morning session met the same lesson groups an equal number of times as did the nine teachers of the afternoon session). While there was a lack of control over "balancing" the lessons for lesson types, the ecological method has the advantage of mitigating against any Hawthorne effect or the even more pervasive problems involved in experimenter effects in general (Orne, 1962; Rosenthal, 1966). This methodology also avoids the different results obtained from "subjects" in experiments as compared to pupils in classrooms, as documented by Kounin (1970). (One might also note that the coders of child behavior did not know the theory that was being tested and, at the time the coding

was taking place, that the researchers didn't either.)

Variables and Their Measurement

With possession of 596 lesson tapes, one investigative problem became the quantification of these phenomena. How could these repeatedly viewable but free-flowing events be translated into quantified variables so that relational questions could be answered?

Dependent Variable

Specification of the dependent variable was straightforward; since lessons were meant to engage or involve children, it became important to determine which varieties of lessons best held the attention of the children and supported the appropriate behaviors. A child's behavior was categorized in one of three ways:

Appropriately involved. This was coded when the child was clearly with the official activity and in a manner appropriate to the thrust of the lesson at that time. The involvement could be passive, for example, listening to a story, watching a demonstration; or it could be active, for example, cutting pictures, pasting objects on cardboard, reciting, singing, dancing.

Not involved. This was coded when the child showed no overt signs of being with the activity yet was not misbehaving or behaving inappropriately.

Inappropriately involved or deviant. Inappropriateness was coded when the child was involved in the activity but in a clearly unsuitable fashion, for example, pounding a magnet against a radiator instead of seeing what sticks to it, "messing" with paste instead of pasting green and red squares on a cardboard, racing about when supposedly skipping to put a valentine in a mailbox. (Making "mistakes" or performing poorly were *not* coded as inappropriate.) A deviant act was an intentional wrongdoing on the part of the child: interference with an ongoing lesson (grabbing the book during teacher's reading a story); aggression against children or property (hitting, pulling hair, throwing equipment against the wall); interfering with legitimate work of other children (taking props away while another was working); open difiance of the teacher.

Each child was coded every six seconds from the time the lesson was started (excluding preparation time) until it was over. Intercoder reliabilities for three coders ranged from 93% to 96% agreement.

The dependent variable, then, was the percentage of off-task behavior in each lesson. This score was computed by dividing the number of six-second units with no involvement, inappropriate behavior, or deviancy by the total number of six-second units coded for that lesson.

Independent Variables

The categorization of the lessons was the major challenge of this research. The ecological niches that lessons represent offer numerous dimensions which might be meaningfully related to pupil involvement. The strategy in this research, first, was to identify lesson *types* rather than single variables. Inferences about operative variables within lesson types were made, but the initial effort was to discriminate lessons as wholes rather than to measure specific unidimensional variables within lessons.

As one observes children in a lesson, it becomes clear that their actions are prodded, oriented, and supported by the external provisions of the lesson. These provisions include the communications of the teacher ("Let's see what sticks and what doesn't stick to a magnet.") and the props that go with the lesson (magnet, paper clips, pieces of paper and cloth, nails). A lesson also includes the standing behavior pattern that goes with the lesson (making piles of objects that stick or don't stick to the magnet, listening to a story being read).

Those external provisions which signal these standard actions can be labelled *signal systems.* One of the principal ways in which lessons differ is in the pattern of their signal systems. Furthermore, lessons' signal systems are like the participation rules of a game: they are repetitive. One can assume that when lesson type X appears a second time, signal system X^1 will also reappear; that is, the external provisions supportive of standard participant action which appeared in the first instance will reoccur in the second.

We may hypothesize that the more continuous and unlagging the provisions of a lesson, the greater the task involvement of a group of children. We may assume that all lessons are planned to produce such continuity and related task involvement. Lessons simply aren't programmed to have nothing to do with nothing. The problem then becomes one of seeing whether certain types of lesson formats are more likely to produce a continuous flow of appropriate signals than are others. A reciprocal problem is to ascertain whether certain formats increase the likelihood of lags or lacunae that may reduce task involvement and make the lesson more vulnerable to inappropriate or deviant behavior.

Let us illustrate the concept of continuity. Common lessons in our data showed teachers reading books or playing records to encircling children. In terms of the signal system concept, this format calls for a single, continuous source of signal emission. This format can be manned by one central occupant as a signal source. Providing this central occupant is adequate, this format produces a continuing signal flow to child participants. This format should yield high involvement or low off-task behaviors.

A single, continuous emitter is also the signal system when a teacher conducts a demonstration, for example, shows how popcorn is made. This latter format also contains persisting external props which invite continuous watching. Again, low off-task behavior should result.

In contrast to the single, continuous source is

a signal system that relies upon multiple, shifting signal sources manned by other children. Such multiple, shifting signal sources are present in formats calling for group discussions (general talk about community helpers), group projects (making a group mural on a felt board), or unrehearsed role play (playing passengers and driver in bus play). In these formats, a common element is that children provide major signals for one another. When these signals are inadequate (as is likely for the performances of unrehearsed children), such a format is vulnerable to lags and faltering continuity. Relatively high off-task behavior should result.

In line with this theory, the more familiar recitation format is similar to the above formats in that it also calls for multiple, shifting signal sources. Each shift in the signal source means that a failure of signal is possible, and when sources are not adequate, the shifts may result in breaks in the signal system. Pupil off-task behavior in these lessons could be predicted to be relatively high.

The success of a formal lesson with a group of children is related to *two* issues: one has to do with the delivery of signals which support appropriate behavior and the other has to do with the prevention of inputs that may encourage inappropriate behavior. We now turn to the discussion of two lesson formats which relate to the latter issue.

Let us consider the case of an individual construction lesson. The teacher provides each child with scissors, paste, a sheet of paper, and magazine pages showing pictures of food and suggests that each child make a collage of desserts. After a child begins such an activity, the major and persisting external signals come from the changing conditions of his materials. He selects a picture, but it must be cut from the page; once cut, the picture requires paste; when paste is applied, it needs to be pressed onto the paper; the remaining space on the paper and the pages of pictures signal selecting another desert, and so on. A continuous signal system occurs as one action and its immediate result provide impetus and guidance for the next action. This signal system and all individual-construction lessons thus provide continuous signals; they should induce high involvement provided each child has appropriate materials and is capable of grasping the goal and of carrying out the necessary participatory actions. It occurred to us, however, that these individual-construction lessons contain a dimension other than continuity which makes it different from the other types of high-continuity lessons. The signal source here, resting as it does on the results of each child's own actions on his own materials, produces a tight, closed behavior–environment circuit. This closed circuit insulates the lesson and shields each child from foreign inputs (distractions, other children's deviancies) which may serve as stimuli to inappropriate behavior. Such a format, in addition to continuity, contains a high degree of *insulation*. This lesson type should produce very low nontask scores.

Another lesson type requiring special consideration is the rather common music-and-movement format. These lessons provide continuous signals from a single source—records or teachers. However, these also contain props (drums, bells) or constituent actions (dancing, jumping, singing) which produce intense stimuli. Since the children are not in a closed-circuit system (such as in individual construction), these stimuli are likely to intrude into participants' attention; that is, children begin to key off one another as well as off the central signal. Such lessons contain high *intrusiveness*. This makes these lessons vulnerable to the spread of inappropriate or deviant behavior, thus increasing the likelihood of off-task scores relative to the other single-emitter systems without intrusiveness.

A signal system code was developed employing the ideas just outlined; all 596 lessons were categorized as belonging to one of the six lesson types (coder agreement 90%). The types, in abbreviated form, are presented in Table 1. The question the research effort asked was whether pupil behavior, specifically off-task behavior, had a statistically significant and psychologically meaningful relation to lesson types as these were described in signal system terms.

Results

Since the investigators exercised no control over which teachers taught which lessons, data were not evenly balanced as they might have been in an experiment. Teachers did present a range of lessons: 33 of the 36 teachers taught at least five of the six lesson types. To maintain constant teachers and groups, a series of t tests for related means was utilized. For example, all teachers and groups who operated in at least one individual-construction lesson *and* in at least one teacher (or record) story lesson were compared. If a teacher taught several lessons in any one type, her score was the mean off-task score for that type. Since not all teachers taught all six types of lessons, the Ns and the means will vary slightly for each comparison.

Table 1 displays lesson types and associated data for off-task percentages. The means in the column labelled $\bar{X}$ are for all teachers who taught those lessons; however, the means in Columns 2 through 6 represent means for teachers who taught *both* of the compared lesson types.

In terms of success in inducing and supporting appropriate task involvement, the lesson types can be grouped into three levels. The best was Type 1; it produced significantly less off-task behavior than all other types. Such individual-construction

TABLE 1

MEAN OFF-TASK SCORES AND DIFFERENCES BETWEEN THE MEAN SCORES FOR THE VARIOUS TYPES OF LESSONS

Lesson type	n	$\bar{X}$	2	3	4	5	6
1. Signals from effects of own behavior on continuously present materials (making individual constructions).	35	5.48	$p < .01$ $n = 32$ $\bar{X}$ 1 = 5.68 $\bar{X}$ 2 = 9.94	$p < .007$ $n = 33$ $\bar{X}$ 1 = 5.50 $\bar{X}$ 3 = 9.32	$p < .001$ $n = 34$ $\bar{X}$ 1 = 5.63 $\bar{X}$ 4 = 11.97	$p < .001$ $n = 29$ $\bar{X}$ 1 = 5.46 $\bar{X}$ 5 = 14.10	$p < .001$ $n = 33$ $\bar{X}$ 1 = 5.46 $\bar{X}$ 6 = 12.89
2. Sequenced signals from a single, continuously emitting source (listening to teacher or records).	33	9.85		$p < .65$ $n = 31$ $\bar{X}$ 2 = 9.97 $\bar{X}$ 3 = 9.07	$p < .19$ $n = 33$ $\bar{X}$ 2 = 9.85 $\bar{X}$ 4 = 12.63	$p < .04$ $n = 27$ $\bar{X}$ 2 = 9.61 $\bar{X}$ 5 = 15.15	$p < .02$ $n = 31$ $\bar{X}$ 2 = 9.47 $\bar{X}$ 6 = 13.63
3. Teacher paces signals to children; also uses continuous external signal source (exploring, reacting to demonstrations).	34	9.25			$p < .05$ $n = 33$ $\bar{X}$ 3 = 9.03 $\bar{X}$ 4 = 11.97	$p < .01$ $n = 29$ $\bar{X}$ 3 = 9.07 $\bar{X}$ 5 = 14.50	$p < .006$ $n = 32$ $\bar{X}$ 3 = 9.12 $\bar{X}$ 6 = 13.02
4. Recitation: with discrete, multiple child signals (dealing with concepts, categories, numbers).	35	12.29				$p < .11$ $n = 29$ $\bar{X}$ 4 = 11.27 $\bar{X}$ 5 = 14.63	$p < .59$ $n = 33$ $\bar{X}$ 4 = 12.29 $\bar{X}$ 6 = 13.24
5. Multiple and shifting signals primarily from child sources (role playing, general talking, group constructing).	30	14.40					$p < .32$ $n = 28$ $\bar{X}$ 5 = 14.41 $\bar{X}$ 6 = 12.72
6. Signals from central source and inputs from high intensity props or actions (singing, moving body).	34	13.05					

Note. The means in the column labelled $\bar{X}$ are for all teachers who taught those lessons; however, the means in Columns 2 through 6 represent means for teachers who taught *both* of the compared lesson types. All p levels are based on two-tailed tests.

lessons contained continuous signals from the results of each child's own action as well as a postulated insulation from external intrusions, apparently a very effective arrangement for this population.

The lessons with average success are the Type 2 and Type 3 formats. These are inferior to the individual-construction lessons, superior to the Type 4, 5, and 6 lessons, and not different from one another. Both Type 2 and Type 3 possess a postulated high degree of continuity and freedom from gaps: this is true of books and records because of sequenced signals from a continually emitting source and it is true of teacher-led demonstrations because of clear sequencing and continually present central props on which there is a continuous focus. It would be reasonable to combine these into one type. When this is done, the combined type is less successful than Type 1 ($p < .001$) and more successful than Type 4 ($p < .023$), Type 5 ($p < .005$), and Type 6 ($p < .001$).

The least successful lessons are Types 4, 5, and 6. These produce significantly more nontask behavior than all the others and are not significantly different from one another. Recitations, role play, and *group* construction have lacunae due to the absence of continuous sequencing and/or to their dependence upon potentially faltering inputs from other children. Even though Type 6 lessons, movement and music, possess a single, continuously emitting source, they are vulnerable to inappropriate behavior because the intense props or behaviors in the format are potentially intrusive.

The model presented here, then, suggests that the pattern and quality of the signal system is a crucial area in predicting child involvement in prescribed lessons in preschool. Three dimensions related to signal input which are suggested by the present research include (*a*) *continuity* of signal input, (*b*) *insulation* of participants from potentially distracting stimuli, and (*c*) *intrusiveness* of respondent action.

Discussion

The following paragraphs compare the explanatory power and empirical validity of this theory with some alternative explanations for the results of this study.

A prevalent theory to account for children's involvement in activities is that of a simple "pleasure" theory: "children enjoy making things" and "they like having stories read to them." Insofar as these particular lessons are concerned, we can only point out that most of the children looked forward to all of the lessons and expected to enjoy all of them. One of our codes was designed to categorize the reaction of the children to a lesson at the point of their first knowing what it was to be about. The reaction of each child at the time the lesson was just started was categorized (with 90% intercoder reliability for a five-category code). Out of 3,002 codable, initial child reactions, 92% were "enthusiastic" or "positive and attentive," 3% were "partly attentive," and 4% were "inattentive" or "negative." It is not surprising, then, that minor variations in initial expectations of pleasure yielded no significant relationship between various measures of initial attraction and involvement.

There is one measure of the pleasure manifested by the children in the actual lessons. Sherman (1971) analyzed all events of "group glee" in the lessons. This was evidenced by one half or more of the group manifesting glee by jumping merrily, screaming joyfully, or laughing. These gleeful incidents occurred most frequently in the motor-expression lessons (Type 6). It is to be noted that these lessons were below average in task involvement. The evidenced pleasure in these lessons does not mean that the lesson format induced the children to remain with the proper behavior requirement nor that it contained protections from inappropriate or deviant behavior.

While one might argue that signal system variations are related to task involvement only when some level of attraction is present, the data do not support the theory that the differences obtained among the lesson types are due to differences in the attraction of the activities.

Other theories of activity have to do with the kind of lesson activity involved. We might point out that in our preliminary

coding of activities, we coded phenotypical qualities of activities and found significant relationships between these and the child behavior scores. Among 21 categories of activities, we found that lessons dealing with construction, stories, or information books had above-average involvement scores; lessons dealing with concepts, discrimination, or categorizations had average scores; lessons dealing with kinesthetics, role play, or general talk had below-average scores. In a code containing eight categories of behavior requirements, we found that lessons in which children constructed or manipulated props had above-average involvement scores; lessons in which children listened or monitored had average scores; and lessons in which children narrated, engaged in bodily movements, or imagined had below-average scores. Although these phenotypical codes possessed predictive power, they did not satisfy us but, instead, tended to be redundant as well as inconsistent with some of the findings. "Construction" lessons, for example, were high-involvement lessons *if* they were individual-construction lessons but were low-involvement lessons if they were group-construction lessons. Listening or monitoring as a behavior mode was associated with high-involvement lessons if the source was a book or teacher, while listening was associated with low involvement if other children were the emitters. Similar inconsistencies were uncovered when we categorized behavior modes as active, passive, or semiactive. Listening or monitoring were categorized as passive yet were associated with different degrees of involvement dependent on what was being monitored. Prop manipulation might be categorized as semiactive since it entailed children "having something to do with their hands," yet these types of behaviors were associated with high involvement when in an individual-construction format and with low involvement when in a context of group construction or playing noisy instruments. While one might descriptively categorize the lesson activities on the basis of phenotypical content and behavior requirements, we feel that the postulated signal system model presents a more genotypical and explanatory theory and resolves some of the inconsistencies noted.

The Sherman (1971) study of group glee may be referred to as adding support to the concepts of insulation and intrusiveness. Group glee, which occurs either in simultaneous bursts or by spreading in a chainlike fashion from one child to another, does provide an indication of the amount of child–child interdependence occurring in a lesson. In our theoretical analysis, we postulated that the individual-construction format is a closed-circuit system in which a child is relatively insulated from inputs from other children. We also theorized that the music–movement formats possessed potential intrusiveness in which a child's intense actions would be likely to intrude into another child's field of attention. The prevalence of group glee provides a measure as to whether this, in fact, does happen in these lessons. Table 2 shows that the rate of glee in the individual-construction lessons is significantly lower than in all other lessons, supporting the theory of its highest insulation. The rate of glee in the music–movement formats is significantly higher than in all other lessons, supporting the theory of its highest intrusiveness.

We deliberately made no attempt, in this preliminary analysis of lesson activities, to measure variations in the dimensions of continuity, insulation, or intrusiveness within any input system. Rather, we are theorizing that one can predict to task-related behavior of a group in a lesson to some degree by a simple knowledge of the skeletal format of a lesson. The results support this contention, even though variations exist among the lessons within any one lesson type.

Conclusions and Implications

The finding that properties of activity settings function to mold the behavior of occupants should have implications for research and practice pertaining to teaching. Perhaps we should devote less effort to the pursuit of teacher characteristics and concentrate more on what techniques are required to program and manage different kinds of activity settings. Different kinds of

TABLE 2

MEAN RATE OF GLEE PER HOUR AND DIFFERENCES BETWEEN THE MEAN SCORES FOR THE VARIOUS TYPES OF LESSONS

Lesson type	2	3	4	5	6
1. Signals from effects of own behavior on continuously present materials (making individual constructions).	$p < .005$ $n = 32$ $\bar{X}\,1 = .697$ $\bar{X}\,2 = 3.502$	$p < .025$ $n = 33$ $\bar{X}\,1 = 1.217$ $\bar{X}\,3 = 3.393$	$p < .005$ $n = 34$ $\bar{X}\,1 = 1.181$ $\bar{X}\,4 = 5.320$	$p < .005$ $n = 29$ $\bar{X}\,1 = 1.279$ $\bar{X}\,5 = 5.743$	$p < .005$ $n = 33$ $\bar{X}\,1 = 1.217$ $\bar{X}\,6 = 11.971$
2. Sequenced signals from a single, continuously emitting source (listening to teacher or records).		$p = ns$ $n = 31$ $\bar{X}\,2 = 3.573$ $\bar{X}\,3 = 3.564$	$p = ns$ $n = 33$ $\bar{X}\,2 = 3.455$ $\bar{X}\,4 = 5.565$	$p = ns$ $n = 27$ $\bar{X}\,2 = 2.931$ $\bar{X}\,5 = 6.177$	$p < .005$ $n = 31$ $\bar{X}\,2 = 3.418$ $\bar{X}\,6 = 11.938$
3. Teacher paces signals to children; also uses continuous, external signal source (exploring, reacting to demonstrations).			$p = ns$ $n = 33$ $\bar{X}\,3 = 3.348$ $\bar{X}\,4 = 5.381$	$p = ns$ $n = 29$ $\bar{X}\,3 = 2.965$ $\bar{X}\,5 = 5.285$	$p < .005$ $n = 32$ $\bar{X}\,3 = 3.200$ $\bar{X}\,6 = 11.686$
4. Recitation: with discrete, multiple child signals (dealing with concepts, categories, numbers).				$p = ns$ $n = 29$ $\bar{X}\,4 = 4.551$ $\bar{X}\,5 = 6.312$	$p < .005$ $n = 33$ $\bar{X}\,4 = 5.214$ $\bar{X}\,6 = 12.327$
5. Multiple, shifting signals primarily from child sources (role playing, general talking, group constructing).					$p < .025$ $n = 28$ $\bar{X}\,5 = 6.418$ $\bar{X}\,6 = 11.464$
6. Signals from central source and inputs from high intensity props or actions (singing, moving body).					

Note. The means in Columns 2 through 6 represent means for teachers who taught both of the compared lesson types. All p levels (one-tailed tests) are based on the Wilcoxon matched-pairs signed-ranks test.

techniques may be required for different kinds of activity settings.

This viewpoint is supported by the studies of Gump (1969) which show differences in teacher and child behaviors dependent upon differences in classroom activity segments and of Kounin (1970) which demonstrate differences in the effectiveness of various teachers' managerial techniques dependent upon whether they are conducting seatwork or recitation settings.

One might note that the concept of continuity of signals may well be related to some dimensions of teaching techniques that Kounin found related to the task involvement of elementary school children. Specifically, smoothness (non-jerkiness) and momentum (absence of slowdowns) were significantly correlated with task involvement in recitation formats. Both momentum and smoothness may now be interpreted as providing lag-free continuity of appropriate lesson inputs.

Preliminary data strongly suggest that lesson formats are clearly related to dimensions of teaching behavior. For example, formats with continuous central emitters require teaching actions somewhat distinct from formats in which there is discontinuous central emission; insulated formats necessitate different actions than formats with intrusiveness. Current research is now ex-

amining the effectiveness of specific teacher action in relation to the general format in which it occurs.

REFERENCES

Barker, R. G. *Ecological psychology: Concepts and methods for studying the environments of human behavior*. Stanford, Calif.: Stanford University Press, 1968.

Gump, P. V. Intra-setting analysis: The third grade classroom as a special but instructive case. In E. P. Willems and H. L. Raush (Eds.), *Naturalistic viewpoints in psychological research*. New York: Holt, Rinehart & Winston, 1969.

Kounin, J. S. *Discipline and group management in classrooms*. New York: Holt, Rinehart & Winston, 1970.

Moos, R. H. Conceptualizations of human environments. *American Psychologist*, 1973, **28**, 652–665.

Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 1962, **17**, 776–784.

Rosenthal, R. *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts, 1966.

Sherman, L. W. Glee in small groups of preschool children. Doctoral dissertation, Wayne State University, 1971.

Shure, M. B. Psychological ecology of a nursery school. *Child Development*, 1963, **34**, 979–992.

(Received September 17, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 563-568

INTELLIGENCE AND TRANSFER: APTITUDE BY TREATMENT INTERACTIONS[1]

GRAHAM R. SKANES,[2] ARTHUR M. SULLIVAN, EDWARD J. ROWE, AND ELIZABETH SHANNON

Institute for Research in Human Abilities, Memorial University of Newfoundland, Canada

The subjects' ability to learn to solve letter-series problems was investigated with low-transfer versus high-transfer practice and in pretest versus no-pretest conditions. Aptitude by treatment interactions were found with Otis IQ and the use of a pretest. Type of practice interacted with Raven IQ. The results were interpreted in terms of Ferguson's, Cattell's, and Jensen's models of the relationship between intelligence and learning. Low-IQ subjects do benefit from direct practice on the strategies involved in the solution of complex problems.

Ferguson (1954, 1956) has proposed that general intelligence is correlated with positive transfer such that subjects with high scores on intelligence tests are expected both to profit more from practice on an intellectual task and to show greater improvement on a related task than low-scoring subjects. This general prediction was confirmed by Sullivan and Skanes (1971), who found that high-scoring (bright) subjects had higher transfer scores than low-scoring (dull) subjects of similar mental age on a letter-series reasoning task following relevant practice. They also noted that the performance of bright subjects was maximized when a pretest was given before the practice session, while dull subjects exhibited their best performance without the pretest, suggesting the presence of an aptitude by treatment interaction (ATI) or an aptitude by instruction interaction (AII).[3] Cronbach and Snow (1969) state that "An ATI exists, in effect, when the regression outcome under treatment A, upon certain pretreatment information, differs in slope from the regression of the same variables under treatment B [p. 4]." In other words, if the functional relationship between learning and intelligence differs under different treatment conditions, this constitutes an aptitude by treatment interaction. In Sullivan and Skanes' (1971) case, it is impossible to speak of regression slopes or functional relationships, since only two extreme groups on the pretreatment variable were used. Nevertheless, when one group achieves optimum learning with one treatment and a second group, differing on intelligence, reveals a maximum with the other treatment, it is reasonable to suspect the existence of different slopes.

The purpose of this study was to determine if different treatments lead to different outcomes for subjects varying on intelligence, as measured by standard intelligence tests. The different treatments were created in two ways: (*a*) by using different types of practice material, one directly related to the learning task (high transfer) and the other less so (low transfer), and (*b*) by using different procedures, one with a pretest condition and one without (cf. Sullivan & Skanes, 1971).

METHOD

Subjects

The subjects were 2,097 school children from Grades 5 to 9 in 13 schools in Corner Brook, New-

[1] This work was done under Grant D-6 from the National Research Council of Canada. Thanks are due Marilyn Tuck who adapted Dowaliby and Berliner's (1971) ANALATI program for our use.

[2] Requests for reprints should be sent to Graham R. Skanes, Department of Psychology, Memorial University of Newfoundland, St. John's, Newfoundland, Canada AIC 5S7.

[3] ATI is the term used by Cronbach and Snow (1969), while Jensen (1970) used AII. Recent reviews of the ATI literature have been provided by Cronbach and Snow (1969) and Bracht (1970a, 1970b).

foundland. They ranged in age from 9 years 9 months to 18 years 8 months.

Materials

Intelligence tests. The Raven Progressive Matrices (Raven, 1938) and the Otis Quick-Scoring Mental Ability Test, Form CM (Otis, 1954) were administered to all subjects.

Learning tests. The learning tests consisted of two forms (A and B) of a letter-series test similar to that in the Reasoning subtest of the Tests of Primary Mental Abilities (Thurstone, 1938). Each item of the 20-item test consisted of a series of letters which formed a logical sequence (see Sullivan & Skanes, 1971). The subject was required to figure out the sequences and to provide the next two letters in the series for each item, thus giving a maximum possible score of 40.

Practice. Two forms of practice were used. The first (number or indirect practice) consisted of 40 items, each of which was a number-series problem (i.e., a series of numbers arranged in a logical sequence). The second (letter or direct practice) consisted of 20 alternating number-series and letter-series problems such that the odd-numbered items were number series and the even-numbered items were letter series. The odd-numbered items on both the number practice and the letter practice were identical, and the even-numbered problems on the number practice were equivalent to the even-numbered problems on the letter practice in terms of sequence presented and method of solution.

Procedure

Both the Raven Progressive Matrices and the Otis Quick-Scoring Mental Ability Test were administered with 30-minute time limits to subjects in their regular classroom groupings. Classrooms (78 in all) were then randomly assigned to the five treatment conditions with the constraint that the number of classrooms from Grades 5, 6, 7, 8, and 9, respectively, be the same in each treatment.

The five treatments were as follows: (*a*) T-T. These subjects were a control group and received Test A with a 10-minute time limit followed 30 minutes later by Test B. (*b*) T-N-T. Test A was administered, followed in order by a 5-minute break, the number-practice session, another 5-minute break, and finally, Test B. (*c*) N-T. These subjects had no pretest. They started with the number practice, had a 5-minute break, and then completed Test B. (*d*) T-L-T. Here the procedure was the same as in condition T-N-T except that the letter practice was substituted. (*e*) L-T. These subjects were treated in the same way as those in condition N-T except that the letter practice was used.

The procedure followed during the practice sessions was the same for the four experimental groups. The experimenter solved the first item, explaining the principle used, and had the subjects write down the correct answer in the spaces provided in the practice form. The subjects then attempted the second problem on their own. This procedure was followed for all of the subsequent items: The experimenter solved and explained the odd-numbered items, and the subjects completed the even-numbered ones. Thus, groups T-N-T and N-T had identical practice sessions, as did groups T-L-T and L-T. Groups T-N-T and N-T differed from groups T-L-T and L-T in that the items completed by the subjects were either number-series items (T-N-T and N-T) or letter-series items (T-L-T and L-T) which paralleled the number-series item just solved by the experimenter. Each practice session took about 20 minutes. Any questions raised by the subjects were answered, so that the session was as much like regular classroom instruction as possible.

Results

Some characteristics of the five treatment groups are given in Table 1. A multiple

TABLE 1

Some Characteristics of the Treatment Groups

Condition	T-T		T-N-T		N-T		T-L-T		L-T	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
Test B	8.1	8.3	12.7	9.6	11.3	7.5	14.2	9.0	13.3	8.2
Age (months)	155.8	22.7	159.7	22.1	157.9	21.2	156.5	20.9	158.8	23.5
Raven IQ	100.0	16.1	100.0	14.6	99.5	13.1	100.9	16.0	102.1	15.8
Otis IQ	98.9	13.6	100.8	13.1	99.7	12.0	101.5	11.8	101.3	11.9
Grade	6.6	1.4	7.1	1.4	6.8	1.3	7.0	1.4	7.0	1.4
% Females	.52		.51		.36		.56		.46	
n	406		409		426		413		443	

Note. Abbreviations: T-T = Test A – Test B; T-N-T = Test A – number practice – Test B; N-T = number practice – Test B; T-L-T = Test A – letter practice – Test B; and L-T = letter practice – Test B.

TABLE 2

MULTIPLE REGRESSION ANALYSIS CONTROL VERSUS PRACTICE GROUPS USING TEST B AS THE DEPENDENT VARIABLE

Variable	Reduction in R^2[a]	F[b]
Practice – no practice (P)	.0267	111.39**
Grade (G)	.0184	76.57**
Sex (S)	.0127	52.93**
Age (A)	.0010	4.04*
Otis IQ (O)	.0388	161.25**
Raven IQ (R)	.0682	283.92**
Interactions (P × G, P × S, P × A, P × O, & P × R)	.0015	1.26

Note. n = 2,097.
[a] R^2 (full model) = .4994.
[b] df = 1/2,086 in all cases except Interactions, where df = 5/2,086.
* $p < .05$.
** $p < .01$.

regression analysis (Cohen, 1968; Overall & Spiegel, 1969) was performed to assess the effects of the practice using Test B as the dependent variable; that is, the control group (T–T) was compared with all the practice groups combined. The full model consisted of the main effects (practice versus no practice, grade, sex, age, Otis IQ, and Raven IQ) and the interactions of the practice–no-practice variable with the other main effects. The results are given in Table 2. All main effects accounted for significant portions of the Test B variance. The proportion of variance predicted by all the interactions combined was not significant. Therefore, the practice had the overall effect of increasing Test B performance, but it did not do so differentially along any of the other main variables, that is, grade, sex, age, Otis IQ, or Raven IQ.

When the control group is excluded, the practice groups form a 2 × 2 factorial design with pretest – no pretest as one factor and type of practice (number versus letter) as the second. A multiple regression analysis (Overall & Spiegel, 1969, Method 2) was performed.[4] The full model included all the main effects and the interactions of the two treatment factors with each of the other main effects. Table 3 gives the results of the analysis. There was no overall pretest effect. All the other main effects were significant predictors of Test B score. The two interactions listed in Table 3 were the only two second-order interactions involving the two treatment factors which were significant. The pretest–no-pretest variable interacted significantly with Otis IQ, and type of practice interacted with Raven IQ.

The two significant interactions were further analyzed to see at what value of the predictor variable significant differences would be obtained between the treatments. The Johnson-Neyman method was used (Johnson & Neyman, 1936; Walker & Lev, 1953) in a computer program written by Dowaliby and Berliner (1971). That program includes the Potthoff criterion (Potthoff, 1964), a more conservative test of the interactions than the orginal Johnson-Neyman criterion (Cahen & Linn, 1971).

The regression lines for the Pretest × Otis IQ interaction are given in Figure 1. The values of Otis IQ, at which significant differences between the pretest and no-pretest conditions occur, are given. The proportion of cases in each region of the graph is also given. Therefore, at the .05 level of significance, children whose Otis IQs are above 100 (about 75% of the cases) do significantly better with the pretest than without.

TABLE 3

MULTIPLE REGRESSION ANALYSIS OF TREATMENT GROUPS USING TEST B AS THE DEPENDENT VARIABLE

Variable	Reduction in R^2[a]	F[b]
Pretest – no pretest (P)	.0011	3.67
Type of practice (T)	.0036	12.18*
Grade	.0163	55.58*
Sex	.0132	44.98*
Age	.0020	6.93*
Otis IQ (O)	.0489	166.74*
Raven IQ (R)	.0781	265.95*
P × O	.0022	7.65*
T × R	.0021	7.27*

Note. n = 1,691.
[a] R^2 (full model) = .5089.
[b] df = 1/1673.
* $p < .01$.

[4] Correlation matrices for each grade level may be obtained from the first author on request.

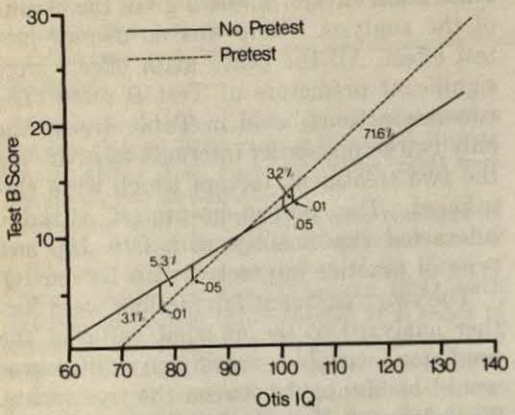


FIGURE 1. The Otis IQ × Pretest–No-Pretest interaction.

The reverse is true for children whose IQs are less than 84 (about 8% of the cases).

The regression lines making up the Raven × Type of Practice interaction are given in Figure 2. At the .05 level, for children with Raven IQs less than 104 (60% of the cases), the direct practice yielded higher Test B scores; while for children with Raven IQs above 133 (about 1% of the cases), the indirect practice proved superior.

DISCUSSION

The experiment reported here has demonstrated aptitude by treatment interactions using standard intelligence tests as aptitude measures. Low transfer versus high transfer and the use of a pretest as opposed to no pretest were the treatments. Cronbach and Snow (1969) have suggested that broad measures of aptitude such as IQ would be the best place to look for aptitude by treatment interactions, and the results obtained here support their view.

The use of a pretest. Campbell and Stanley (1963) have suggested several possible effects of pretesting in experimental methodology. The usual effect is that pretesting leads to improved performance on retest. The results reported here demonstrate that the effect of a pretest is a function of the intelligence of the child. Therefore, it is possible that in some experiments a pretest might obscure treatment effects in subjects with certain characteristics. The indiscriminant use of pretests in experiments might obscure possible treatment effects because the pretest interacts with the treatment to produce a decrement in performance.

Samuels (1969) and Hartley (1971) have, however, supported the view that a pretest enhances performance after practice. Samuels found that pretests with feedback resulted in greater retention of the material read. Hartley, using programmed instruction, found that the pretest has no appreciable effect if learning is efficient but that it facilitates performance when the task is difficult. Sullivan and Skanes (1971) added that the pretest effect is a function of the ability of the learner. The aptitude by treatment interaction reported here strengthens that view because the pretest effect is found to be true when the practice used is direct or indirect. The effect appears when aptitude is measured by the Otis IQ test but not when it is measured by the Raven. For high-Otis-IQ subjects, the pretest resulted in higher performances than no pretest. For low-Otis-IQ subjects, the no-pretest condition was superior to the pretest condition.

Little can be said for certain here about the cause of the pretest effect. It might be motivational, with lower-IQ subjects becoming discouraged by the difficult and probably meaningless problems with which they are faced. Hartley (1971) preferred an explanation in terms of tuning. He suggested that the pretest will "not only . . . alter one's expectations to what is required, but . . .

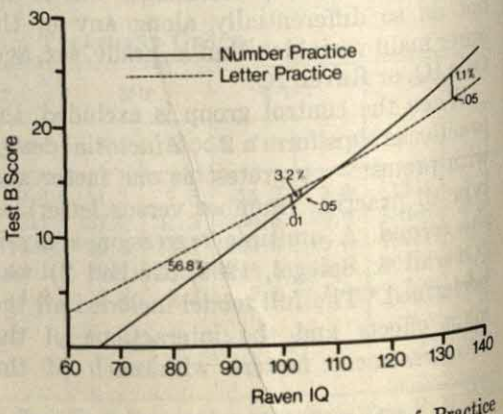


FIGURE 2. The Raven IQ × Type of Practice interaction.

seems to assist in the organization of other related material so that it is easily remembered [p. 143]."

The task used here appears to require of the subject that he formulate hypotheses about the solution of the problem encountered (e.g., "two ascending series interposed"). He then has to check the hypothesis against the material in view and, if it fits, extend it. That process is akin to "formal operational reasoning," the highest stage of intellectual development in Piaget's system. The pretest gives subjects an opportunity to see what the problem is and what possible solutions there are. High-IQ subjects can formulate and test hypotheses during the pretest; low-IQ subjects probably cannot. The practice session for high-IQ subjects is a session in which already formulated notions about strategies are honed and sharpened, incorrect ones rejected, and new leads formed. The practice session for the low-IQ subjects is imposed on a muddled background and probably serves to consolidate confusion.

Differential transfer conditions. It is interesting from the point of view of Ferguson's (1954, 1956) theory of transfer and intelligence that the Raven IQ, generally agreed to be the best single measure of general intelligence in Spearman's sense, yielded an interaction with the different transfer treatments but that the Otis IQ did not. Ferguson has suggested that general intelligence is positively related to the ability to transfer. The fact that the regression line relating learning to Raven IQ in the indirect or low-transfer condition was significantly steeper than in the direct or high-transfer case indicates that Ferguson is correct. Subjects whose level of general intelligence was high used the difficult transfer situation to advantage, while the low-IQ subjects were unable to do so and profited more from the direct training.

The aptitude by treatment interactions found in this study are important in view of the current controversy concerning the relevance of genetic or constitutional factors in IQ and learning. Two related but separate propositions have been espoused: first, by Cattell and his associates (Cattell, 1963; Horn, 1968) and, second, by Jensen (1969a, 1969b, 1970).

In the Cattell (1963) model, there are two types of intelligence, fluid and crystallized. The former is much like ability to transfer, while the latter resembles the ability to acquire information. Horn (1968), in reviewing the characteristics of various tasks, has noted that the Raven is a good measure of fluid intelligence and that such series tests as the ones used here for the learning task are also good measures of fluid ability. Fluid ability in the Cattell view is highly constitutional in nature, and in Horn's scheme, it is related to neurological functioning.

Our results can be interpreted in the following way using Cattell's model. The direct practice (high-transfer) condition is analogous to a situation in which crystallized ability is useful. The indirect practice (low-transfer) condition is one in which fluid intelligence would be needed. The slope of the regression line in the crystallized case was flatter than in the fluid, confirming the idea that the low-transfer situation was more directly related to an accepted measure of fluid intelligence than was the other.

In a similar way, Jensen's (1969a, 1969b, 1970) model can be used as a framework onto which the results can be fitted. Jensen has proposed that some complex tasks (tasks requiring reasoning as opposed to rote learning) are relatively free of cultural factors and, hence, are highly heritable. The Raven Progressive Matrices is one such task. We can conclude, in Jensen's terms, then, that the direct practice condition is not as highly dependent upon inherited ability as the indirect practice condition because the former is less directly related to Raven IQ.

The data, therefore, support the Cattell and Jensen views. But more important, they show a way around the conclusions that Cattell and Jensen appear to formulate; namely, that people of low fluid intelligence or low, culture-free, complex learning ability might be unable, because the factors are inherited or constitutional, to do complex tasks. Our data show that when a complex task is presented in a direct manner, sub-

stantial improvement in performance is possible.

Guinagh (1971) has done something similar. He trained children on the strategies relevant to solving the problems in the Raven Progressive Matrices. Even groups whose performance would be predicted to remain low because of presumed genetic impoverishment, showed substantial improvement. In other words, direct training on the strategies to be employed to do complex tasks leads to increased performance by subjects who would not be expected to perform well. Or, strategies can be learned by rote and applied to problems in much the same way as if the strategies were derived by the subject.

The conclusion from this study and from Guinagh's (1971) study is that the effort to make tests culture fair by choosing novel material (e.g., Raven's matrices) or completely familiar material (e.g., the alphabet; Horn, 1968) did not succeed because test makers neglected to ensure that the relevant strategies were equally novel or equally familiar. While the evidence presented here in no way suggests an elimination of intelligence factors in learning, it does point the way to a modification of the extreme interpretation of Jensen's work.

REFERENCES

Bracht, G. T. Experimental factors related to aptitude-treatment interactions. *Review of Educational Research*, 1970, **40**, 627–645. (a)

Bracht, G. T. *The relationship of treatment tasks, personological variables, and dependent variables to aptitude-treatment interactions.* (Doctoral dissertation, University of Colorado), Ann Arbor, Mich.: University Microfilms, 1970, No. 70-5820. (b)

Cahen, B. S., & Linn, R. L. Regions of significant criterion differences in aptitude-treatment-interaction research. *American Educational Research Journal*, 1971, **8**, 521–530.

Campbell, D. T., & Stanley, T. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Cattell, R. B. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 1963, **54**, 1–22.

Cohen, J. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 1968, **70**, 426–443.

Cronbach, L. J., & Snow, R. E. *Individual differences in learning ability as a function of instructional variables.* (Final Report, Stanford University, Contract OEC-4-6-061269-1217) Washington, D.C.: U. S. Department of Health, Educational, and Welfare, Office of Education, March 1969.

Dowaliby, F. J., & Berliner, D. C. *ANALATI: A computer program for the application of the Johnson-Neyman technique in the case of one predictor and one criterion variable.* (Tech. Rep. 15) Amherst, Mass: University of Massachusetts School of Education, Center for Educational Research, 1971.

Ferguson, G. A. On learning and human ability. *Canadian Journal of Psychology*, 1954, **8**, 95–112.

Ferguson, G. A. On transfer and the abilities of man. *Canadian Journal of Psychology*, 1956, **10**, 121–131.

Guinagh, B. J. An experimental study of basic learning ability and intelligence in low-socio-economic-status children. *Child Development*, 1971, **42**, 27–36.

Hartley, J. Factors affecting the efficiency of learning from programmed instruction. *AV Communication Review*, 1971, **19**, 133–148.

Horn, J. L. Organization of abilities and the development of intelligence. *Psychological Review*, 1968, **75**, 242–259.

Jensen, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 1969, **39**, 1–123 (a).

Jensen, A. R. Intelligence learning ability and socioeconomic status. *Journal of Special Education*, 1969, **3**, 23–35 (b).

Jensen, A. R. Hierarchical theories of mental ability. In W. B. Dockrell (Ed.), *On intelligence*. Toronto: Ontario Institute for Studies in Education, 1970.

Johnson, P. O., & Neyman, J. Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, 1936, **1**, 57–93.

Otis, A. S. *Otis Quick-Scoring Mental Ability Test.* New York: Harcourt, Brace & World, 1954.

Overall, J. E., & Spiegel, D. K. Concerning least squares analysis of experimental data. *Psychological Bulletin*, 1969, **72**, 311–322.

Potthoff, R. R. On the Johnson-Neyman technique and some extensions thereof. *Psychometrika*, 1964, **29**, 241–256.

Raven, J. C. *Progressive Matrices, Sets A, B, C, D, and E.* London: Lewis, 1938.

Thurstone, L. L. *Primary mental abilities.* Chicago: University of Chicago Press, 1938.

Samuels, S. J. The effect of post-test relevant pre-tests and discussion-type feedback on learning and retention. *Psychonomic Science*, 1969, **16**, 67–68.

Sullivan, A. M., & Skanes, G. R. Differential transfer of training in bright and dull subjects of the same mental age. *British Journal of Educational Psychology*, 1971, **41**, 287–293.

Walker, H. M., & Lev, J. *Inferential statistics.* New York: Holt, 1953.

(Received September 17, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 569-571

COGNITIVE STYLE AND REASONING ABOUT SPEED

LINNEA C. EHRI[1] AND IRENE M. MUZIO

University of California, Davis

College students were given a test for field dependency, and they were asked to reason about the relative speeds of horses turning on a merry-go-round platform. To promote the recognition that horses located on the outside of the platform traveled faster than horses on the inside, additional problems were presented which directed subjects' attention to the relevant variables. Results revealed that unlike field independent subjects who reasoned correctly from the outset, field dependent subjects failed to think analytically. They were misled by perceptually salient aspects of the situation; they tended to center on these features in their reasoning; and they resisted accommodating to additional information.

Variations in cognitive styles have been of interest to investigators as a means of accounting for individual differences in performance on perceptual and conceptual tasks. Not only Witkin and his associates (Witkin, Dyk, Fattuson, Goodenough, & Karp, 1962; Witkin, Lewis, Hertzman, Machover, Meissner, & Wapner, 1954) but also Kagan, Moss, and Sigel (1963) have produced evidence suggesting that two opposing modes of thinking can be identified: (*a*) analytic or field independent, the ability to distinguish, hold in mind, and coordinate the relevant figures or attributes extracted from a more complex and sometimes distracting stimulus context, and (*b*) global or field dependent, a more passive, unanalyzed processing of the stimulus in which its holistic nature is preserved and only the most obvious relations among its parts are manipulated mentally. This distinction has been incorporated in a performance model of cognitive development proposed by Pascual-Leone (1970). He has suggested that Piaget's theory of intellectual functioning be extended to include a central processor or computing space whose function it is to transform and coordinate information and whose capacity expands as the child grows. And he has reported evidence indicating that field dependent subjects, unlike field independent subjects, often fail to utilize the full structural capacities of their central processors. That is, they do not extract and mentally manipulate the maximum number of independent pieces of information which their computing spaces are capable of handling. As a consequence, these subjects display reasoning patterns similar to those of younger children observed by Piaget and Inhelder (1941).

The relevance of cognitive styles and Pascual-Leone's model of information processing became of special interest to the present authors as a means of accounting for some puzzling differences observed in the thinking of college students. In response to the question, Would a child riding on the outside row of horses attached to a moving merry-go-round platform be traveling faster than a child circling on the inside row of horses or would they be moving at the same speed?, some people argued that the outside moved faster, but an equal number insisted that both moved at the same speed, and a few claimed that the inside traveled faster. Analysis of the reasons presented in support of these positions pointed to the possibility that patterns might be reflecting differences in the general style of thinking adopted by adults to solve problems. To determine whether alternative solutions to the moving merry-go-round problem might be related to differences in field dependency among adults, the following study was undertaken.

[1] Requests for reprints should be sent to Linnea C. Ehri, Department of Education, University of California, Davis, California 95616.

Method

Four verbally stated problems were constructed. The first instructed subjects to think of a merry-go-round which has two circles of horses. One circle is on the outside of the platform. The other circle is on the inside. Billy selects a horse on the outside to ride. Danny picks a horse on the inside. The merry-go-round begins turning. Is one boy going faster than the other boy or are they both going the same speed? The subjects were required to check one of the three alternatives (i.e., outside faster, inside faster, same speed) and then to explain their choice. Three problems designed to assist the subjects in focusing upon relevant aspects of the moving merry-go-round problem were subsequently presented. These required subjects to notice differences in distance traveled by tracing the routes of the horses, to think about an analogous situation in which children hold hands and run around a post (i.e., Do some children run faster?), and to consider the formula for speed. Following each of these problems, the subjects were asked to reconsider their answers to the moving merry-go-round problem and to decide whether they still believed their answers to be true or whether they wished to alter their positions.

Field dependence was measured with an embedded figures test, the Components subtest (Flanagan Aptitude Classification Tests, 1958). It consisted of two parts, each part requiring the subjects to identify which of 5 simple geometric figures was embedded in each of 20 very complex figures.

All tasks were printed and administered to groups of students ($n = 61$) enrolled at the University of California, Davis. The subjects provided written answers.

TABLE 1

Results of the Series of Moving Merry-Go-Round Problems

Score on the embedded figures test[a]	Moving merry-go-round problems			
	Solved from start	Solved by end	Never solved	Total
Field dependent (below 20)	1	3	11	15
Middle range (20–30)	9	12	8	29
Field independent (above 30)	11	4	2	17
Total	21	19	21	61

Note. Results indicate the number of field independent and field dependent subjects who concluded or never concluded that the outside child was moving faster in the series of moving merry-go-round problems.

[a] The scores are those obtained on the Composite subtest of the Flanagan Aptitude Classification Test.

Results

Performances on the four problems were quantified. The subjects were divided into three groups: (*a*) those who concluded from from the beginning that the outside child was going faster ($n = 21$); (*b*) those who shifted to this position by the end of the problems ($n = 19$); (*c*) those who never arrived at this conclusion ($n = 21$). A score of 1, 2, or 3 was assigned to each subject according to his group membership, and these values were correlated with scores on the embedded figures test. Results revealed that field dependence was indeed related to performance on the moving merry-go-round problem ($r = -.49, p < .01$). This relationship was verified in another way (see Table 1). A chi-square test revealed that the distributions of field dependent and field independent thinkers in the three groups were distinctly different ($\chi^2 = 14.58$, $df = 2$, $p < .01$). The majority of field independent subjects tended to reason from the beginning that the outside child was traveling faster. In contrast, most field dependent subjects neither proposed this solution initially nor arrived at the "outside" answer by the end of the problem set. Although the additional problems were successful in altering the views of 19 subjects, few of these (only 3 subjects) were classed as field dependent while the majority fell in the middle range on the field dependency measure.

Of the 21 subjects who reasoned correctly from the outset, 17 provided complete explanations citing the greater distance covered in the same amount of time or in one revolution by the outside child. The remainder either mentioned only one of these variables explicitly or reasoned by a crack-the-whip analogy.

Examination of the reasoning patterns of subjects who persisted throughout the problems in arguing that the boys were going the same speed revealed the following justifications: the platform consists of one circle turning at one, not two speeds; the platform is the only thing moving, not the horses which are attached to it; there exists only one belt or engine generating one, not two,

movements. Although these subjects saw the outside child as moving faster in the post problem, they all regarded this as unrelated to the moving merry-go-round problem for reasons which focused upon unanalyzed physical differences in the two situations (i.e., "the children are running, the horses are not"). None of the subjects distinguished between linear and rotational notions of speed to resolve the discrepancy. Such reasoning provides a clear illustration of field dependent thinking in a context which is quite different from the embedded figures test.

Although the majority of the faulty arguments favored "same speed," there were some subjects who mentioned that the inside child might be moving faster. In their explanations, these subjects appeared to forget that the horses remained evenly positioned on the platform. Noticing that the circumferences of the paths differed, they decided that the inside child could not be going slower and might even be circling faster since he had less far to go than the outside child. It was observed, in fact, that the route-tracing problem which focused attention on these differences in distance had the effect of strengthening this line of reasoning in seven cases.

Discussion

Reasoning exposed in the set of moving merry-go-round problems was successful in providing evidence consistent both with cognitive style distinctions and with Pascual-Leone's (1970) information-processing model. Whereas field independent subjects were able to analyze the stimulus contexts, extract the relevant variables, and coordinate them appropriately, the thinking of field dependent subjects tended to be dominated by the perceptible physical properties of the total stimulus configuration, and these subjects were most resistant to the influence of prompts hinting at another line of reasoning. Of course, given that a global, intuitive interpretation lacks clear structuring among parts, it is perhaps not surprising that such a product is less amenable to either slight or substantial modification in the face of additional information.

REFERENCES

Flanagan, J. C. *Flanagan Aptitude Classification Tests.* Chicago: Science Research Associates, 1958.

Kagan, J., Moss, H. A., & Sigel, I. E. Psychological significance of styles of conceptualization. In J. C. Wright & J. Kagan (Eds.), Basic cognitive processes in children. *Monograph of the Society for Research in Child Development,* 1963, **28,** (2, Serial No. 86).

Pascual-Leone, J. A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychologica,* 1970, **32,** 301–345.

Piaget, J., & Inhelder, B. *Le developpement des quantités chez l'enfant.* Neuchâtel, France: Delachaux et Niestlé, 1941.

Witkin, H. A., Dyk, R. B., Fattuson, H. F., Goodenough, D. R., & Karp, S. A. *Psychological differentiation.* New York: Wiley, 1962.

Witkin, H. A., Lewis, H. B., Hertzman, M., Machover, K., Meissner, P. B., & Wapner, S. *Personality through perception: An experimental and clinical study.* New York: Harper, 1954.

(Received September 21, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 572-579

"ENVIRONMENTS" OF BRITISH COMMONWEALTH UNIVERSITIES

JAMES M. RICHARDS, JR.[1]

The Johns Hopkins University

The purpose of this study is to use some techniques developed for describing the "environments" of U.S. universities to explore the correlation between national characteristics and university characteristics in the British Commonwealth. Because such techniques have been little used outside the United States, the study first examines their appropriateness for describing Commonwealth universities. Finally, the study uses one of these techniques to examine the collegial organization of Cambridge University and Oxford University, England. Results for 186 universities suggest that these techniques are appropriate for characterizing Commonwealth universities, including Oxford and Cambridge colleges, and that national characteristics and university environments correlate fairly meaningfully in the Commonwealth. Therefore, this study helps provide a broader, more international context for studying university environments.

Universities have goals and traditions that transcend differences among nations; therefore, an interesting question is whether nations with different characteristics also have universities with different characteristics. Techniques for describing how colleges and universities differ have been developed by psychologists and other behavioral scientists in the United States (Astin, 1962, 1968; Astin & Holland, 1961; Pace, 1963; Richards, Bulkeley, & Richards, 1972; Richards, Seligman, & Jones, 1970; Stern, 1970), but only a few studies (Richards, 1973; Richards, Rand, & Rand, 1968) have used these techniques to study institutions outside the United States. The purpose of the present study is to use some of these techniques to explore the extent to which national characteristics and university characteristics (or university "environments") correlate in the British Commonwealth. Because such techniques have been little used outside the United States, the study first examines evidence of the appropriateness of these techniques for characterizing British Commonwealth universities. Finally, the study uses one of these techniques to examine the collegial organization of Oxford University and Cambridge University, England. Holland's (1973) theory of interpersonal environments provides an overall framework for much of this investigation.

METHOD

Procedure

Sample of universities. The source of data for this study was the *Commonwealth Universities Yearbook 1969* (Association of Commonwealth Universities, 1969). This compendium includes data for member institutions of the Association, plus universities in Ireland and South Africa,[2] and presents detailed information for all of these universities except those established most recently.

The sample consisted of 186 universities with complete data. Whenever the *Yearbook* data permitted, geographically separated campuses with a common administration were treated as separate institutions. For example, the three campuses of the University of the West Indies (in Jamaica, Trinidad, and Barbados) were treated separately.

[1] Requests for reprints should be sent to James M. Richards, Jr., Center for Social Organization of Schools, Johns Hopkins University, Baltimore, Maryland 21218.

[2] Inclusion of universities from a given nation, of course, implies neither approval nor disapproval of the social and political organization of that nation.

The major problems, and perhaps the major biases, for this study occurred in India and Pakistan, where most universities are "affiliative" examining bodies with little if any centralized faculty and instruction. This study includes only those universities in these two countries that appeared to be "unitary teaching" universities, either by official designation or in terms of the data presented in the *Yearbook*. The examining-body form of university organization is modeled on the University of London, England, and that university too was excluded from this study to reduce differential bias among nations. For analyses at the university level, Oxford University and Cambridge University, England, were treated as unitary institutions despite their collegial organization.

It is arguable whether this group of universities constitutes a sample or a population (and therefore whether it is meaningful to use tests of statistical significance). In a strict statistical sense, this group seems more nearly a population, but this study is of less interest if one does not make some sort of generalization beyond these particular universities. Also, some readers may choose to regard these universities as a sample and wish to know the statistical significance of the results. Accordingly, significance levels are reported where computable.

Environmental measures. Two sets of environmental measures were used in this study. The first grows out of Holland's (1973) theory that people in various occupations have characteristic personalities, either because people with differing personalities are attracted to characteristic occupations or because differing occupations mold personality in characteristic ways. Specifically, one's occupation corresponds to one's personality type—realistic, investigative, artistic, social, enterprising, or conventional. It follows that an interpersonal environment is a function of the number, or relative number, of people with these personality types composing that environment. The basic procedure for assessing an environment, therefore, is simply to count the number of people in that environment whose occupations fall into each of the six types. Usually some control for the size of the environment (i.e., the total number of people) is necessary, such as computing the percentage of the people who fall into the six types. Variations of this procedure have been successfully applied to measuring college and university environments in the United States (Astin, 1965; Astin & Holland, 1961; Richards et al., 1970, 1972) and in Japan (Richards, 1973).

In the present study, the number of faculty members at each Commonwealth university falling into each type was determined. To reduce skewness, a square root transformation was used for most analyses involving the number of faculty falling into the six types. Assignment to types was based on the specific academic discipline cited for each individual faculty member in the *Yearbook*. Disciplines were assigned to types on the basis of several empirical classifications for occupations and major fields developed by Holland and his associates (Holland, 1966; Holland, Viernstein, Kuo, Karweit, & Blum, 1970; Viernstein, 1971). *Realistic* fields are typified by agriculture and civil engineering; *investigative* fields by biology and physics; *artistic* fields by literature and foreign languages; *social* fields by education and sociology; *enterprising* fields by economics and law; and *conventional* fields by accounting. The number of faculty in a given type not only measures *absolute* emphasis on that type but also strongly involves university size. The percentage of faculty in each type at each university was also computed to measure *relative* emphasis on the six types. These percentages were transformed separately for each type to normalized standard scores by the percentile procedure. While this transformation did succeed in making the scores more nearly normal, its major effect was to place them on a scale more suitable for correlational analysis. Regardless of the transformation, of course, the percentage scores are ipsative in that a university with a high score on one type must have low scores on the other types. Therefore, statistical tests involving the six distributions of percentages are not completely independent.

The second set of environmental measures used in this study are based on factor-analytic investigations of the environments of two-year (Richards, Rand, & Rand, 1966) and four-year (Astin, 1962) institutions in the United States and of universities in Japan (Richards, 1973). Three factors appeared common to all of these studies. The first was simply institutional *size*. The second was *affluence* in the sense of having a large budget and extensive facilities (e.g., library, faculty, etc.) relative to the number of students. The third might best be termed *technological emphasis*. High-scoring institutions offer extensive training in engineering, have a large percentage of males in their student body, and are characterized by secular (rather than religious) control or orientation. Because comparable results were obtained in the United States and Japan, it appears that these three dimensions may be appropriate for universities in many nations.

Accordingly, measures of size, affluence, and technological emphasis were developed for each university.[3] Size was measured by the total number of full-time students. A square root transformation was made to reduce skewness. The measure of affluence involved two variables: (*a*) university income per full-time student (in Canadian dollars at 1969 exchange rates) and (*b*) the number of library books per full-time student. These variables were equated for mean and standard deviation and combined with unit weight. Tech-

[3] It appeared more efficient, and more appropriate to the *Yearbook* data, to calculate these measures directly rather than first conducting a similar factor-analytic investigation of Commonwealth universities.

nological emphasis was also measured by a unit-weighted combination of two equated variables chosen on the basis of empirical results from previous studies (Astin, 1962; Richards, 1973). To avoid overlap with the *realistic* type score, no direct measure of technological education was included. The first variable was the percentage of males in the full-time student body, and the second was a measure of secular orientation in which (prior to the equating transformation) universities under religious control scored 1, secular universities offering training in religious subjects scored 2, and secular universities without religious training scored 3. Last, the distributions on size, wealth, and technological emphasis were converted separately to (unnormalized) standard scores ($M = 50, SD = 10$).

All of these variables represent nonreactive measures (Webb, Campbell, Schwartz, & Sechrest, 1966) of the "objective" university environment. In the United States, such measures have been shown to be correlated with reactive measures of the psychological, or phenomenological, university environment (Astin, 1963; Richards et al., 1970), but it has not been determined whether this is true in the British Commonwealth. It is also undetermined whether Holland's notions about the relationship between occupation and personality apply in the Commonwealth, although there is some evidence (Lonner, 1968) that vocational interests and some occupations are similarly related in several non-Commonwealth nations. Finally, this study represents a "man from Mars" point of view (Townsend, 1971) in that the investigator has no personal experience with any Commonwealth university. Therefore, this first study of Commonwealth university environments should be interpreted with restraint.

Results

Table 1 summarizes the central tendencies and variabilities of the type scores. A marginal totals chi-square—based on the total number in each type at each university—rejected ($\chi^2 = 14{,}386.92$, $df = 925$) the null hypothesis of no variation among universities with respect to type distributions well beyond the .01 level. Most emphasis is placed on investigative, artistic, and social fields and the least on conventional fields. In other words, Commonwealth universities appear to place most emphasis on the physical and biological sciences, next most on the liberal arts, and next most on such fields as teacher education, the social sciences, etc. This pattern is similar to the one obtained for universities in the United States (Richards et al., 1970) and in Japan (Richards, 1973).

TABLE 1

Measures of Central Tendency and Variability for Faculty Types

Type scores	No. faculty ($\sqrt{n}$)		% Faculty	
	M	*SD*	Median	Semi-inter-quartile range
Realistic	5.15	3.37	7.96	6.29
Investigative	11.29	5.80	42.14	9.67
Artistic	6.87	3.38	17.17	6.42
Social	6.63	5.86	14.86	6.54
Enterprising	4.94	2.83	8.71	3.51
Conventional	1.03	1.22	0.13	0.40

TABLE 2

Correlations of Type Scores with Size, Affluence, and Technological Emphases for All Commonwealth Universities Combined

Type scores	Size	Affluence	Technological emphasis
No. faculty			
Realistic	.58**	−.19**	.33**
Investigative	.84**	.03	−.04
Artistic	.80**	.08	−.38**
Social	.48**	.06	−.29**
Enterprising	.73**	.05	−.26**
Conventional	.52**	−.05	−.23**
% faculty			
Realistic	.11	−.28**	.61**
Investigative	.27**	.01	.33**
Artistic	−.05	.12*	−.57**
Social	−.10	.17**	−.62**
Enterprising	.01	.06	−.29**
Conventional	.28**	−.03	−.22**

Note. $n = 186$.
* $p < .05$.
** $p < .01$.

Next, correlations were computed between the type scores and the measures of size, affluence, and technological emphasis. Table 2 summarizes these results. More than half of the correlations are significant, and each environmental measure is correlated with several type scores.[4] These correlations seem to reflect primarily contrasts between large and small universities and be-

[4] It should be reemphasized that the type measures are ipsative so the significance tests are not entirely independent.

tween universities which do and do not emphasize technological training. The pattern of correlations between technological emphasis and the percentage of faculty in the various types is similar to the pattern of correlations found for universities in the United States (Richards et al., 1970) and in Japan (Richards, 1973). That is, technological emphasis is positively correlated with an emphasis on realistic and investigative fields and negatively correlated with an artistic and social emphasis. The patterns of correlations for size and affluence, however, show considerable deviation from the pattern obtained in these earlier studies. In both the United States and Japan, size is positively correlated with a relative emphasis on realistic fields and is not significantly correlated with an investigative emphasis. (In all countries, of course, size is positively correlated with the *number* of faculty members in each of the types.) In the British Commonwealth, size is positively correlated with a relative emphasis on investigative fields and is not significantly correlated with a realistic emphasis. Similarly, in the United States and Japan, affluence is correlated positively with realis- and negatively with artistic, while the reverse pattern holds in the Commonwealth. These differences presumably can be attributed to differing histories of university development in the various nations.

To check the extent to which national differences in these relationships occur *within* the British Commonwealth, the environmental measures were correlated with the type scores separately for universities in Britain and in Canada (the only nations with enough universities for such computations to be meaningful). Table 3 summarizes the results. Again, the results suggest reasonable consistency across nations for the correlates of technological emphasis but some variation between nations for the correlates of size and affluence. If we interpret these results as population correlations for which statistical tests are irrelevant, several of the correlations involving affluence are of opposite sign for Britain and Canada. This pattern certainly suggests differences in the fields in which these nations invest, but it does not indicate whether such differences reflect value differences or merely reflect historical contingencies.

The next analysis examined the question of whether national characteristics and university characteristics are correlated. Sawyer (1967) found that much of the variation among nations can be described

TABLE 3
CORRELATIONS OF TYPE SCORES WITH SIZE, AFFLUENCE, AND TECHNOLOGICAL EMPHASIS FOR UNIVERSITIES IN INDIVIDUAL NATIONS

Type scores	Britain (n = 50)			Canada (n = 48)		
	Size	Affluence	Technological emphasis	Size	Affluence	Technological emphasis
No. faculty						
Realistic	.72**	−.24	.20	.68**	.05	.36*
Investigative	.91**	−.03	−.06	.94**	−.08	.12
Artistic	.79**	.15	−.57**	.87**	−.18	.01
Social	.37**	−.01	−.28*	.90**	−.19	−.10
Enterprising	.82**	−.11	−.12	.82**	−.08	.04
Conventional	.56**	−.28*	−.21	.55**	−.23	−.01
% faculty						
Realistic	.33*	−.38**	.45**	.20	.18	.72**
Investigative	.28*	−.06	.42**	.54**	.04	.48**
Artistic	−.20	.22	−.60**	−.43**	−.13	−.45**
Social	−.02	.20	−.56**	−.15	−.21	−.75**
Enterprising	.01	−.18	.04	−.11	.22	−.20
Conventional	.41**	−.32*	−.22	.22	−.21	.01

* $p < .05$.
** $p < .01$.

TABLE 4

CORRELATIONS BETWEEN NATIONAL CHARACTERISTICS AND AVERAGE CHARACTERISTICS OF UNIVERSITIES WITHIN NATIONS

Average characteristics of universities	Characteristics of nations		
	Size	Wealth	Political orientation
Environmental measures			
Size	.55**	.43*	−.44*
Affluence	−.26	.38*	.30
Technological emphasis	.16	−.28	.29
No. faculty			
Realistic	.66**	−.01	−.09
Investigative	.52**	.30	−.40*
Artistic	.42*	.59**	−.43*
Social	.28	.12	−.05
Enterprising	.14	.05	−.17
Conventional	.22	.53**	−.49**
% faculty			
Realistic	.44*	−.15	.09
Investigative	.05	.12	−.10
Artistic	−.21	.47**	−.17
Social	−.32	−.16	.41*
Enterprising	−.37	−.26	.14
Conventional	.12	.41*	−.37

Note. $n = 27$.
* $p < .05$.
** $p < .01$.

by just three relatively independent dimensions: *size, wealth,* and *political orientation* (i.e., Communist, neutral or anti-Communist). Measures of these three dimensions were determined for each of 27 nations (or political subdivisions) which included one or more of the studied universities. Size was measured by the total national population, wealth by per capita income (in Canadian dollars at 1969 exchange rates), and political orientation, following Sawyer, by a trichotomous variable with scores of 1 assigned to nations having a military alliance with the United States, scores of 2 to "neutral" nations, and scores of 3 to Communist nations. In the present study, of course, this variable primarily contrasts "neutrality" versus military alliance with the United States.

In some cases, a judgmental element was involved in deciding how to treat a particular university or geographic subdivision. Each campus of the University of the West Indies (Jamaica, Trinidad, and Barbados) again was treated as a separate university, and each island incorporating a campus was treated as a separate nation. A similar procedure was followed for the various campuses of the University of East Africa (Kenya, Tanzania, and Uganda). On the other hand, Lesotho alone was treated as the home of the University of Botswana, Lesotho, and Swaziland, since the only campus is located in Lesotho.

The characteristics of nations were correlated with the *average* university profiles within nations. Results are shown in Table 4. It appears from these results that the characteristics of universities do indeed vary systematically with the characteristics of British Commonwealth nations. Large nations have large universities that emphasize realistic fields; wealthy nations have large, affluent universities that emphasize artistic and conventional fields. (Perhaps a certain amount of wealth is necessary before it is worthwhile educating accountants to keep track of it.) Neutral countries tend to have small universities which emphasize social fields. In most cases, this means they emphasize teacher training.

Some Commonwealth universities, most notably Cambridge University and Oxford University, England, are organized more into relatively autonomous colleges than into a centralized university structure. This organizational pattern is especially suitable for exploring the extent to which Holland's (1973) typology also is useful for measuring subenvironments within universities, so the last analysis for this study used the typology to examine the environments of Oxford and Cambridge colleges. Specifically, a count was made of the number of faculty members at the various colleges falling into Holland's types. (This count included all "members" of these colleges rather than just "fellows.") Because only one individual was classified as conventional, that type was excluded from subsequent analyses for these colleges.

A square root transformation was used to reduce skewness. The transformed data for colleges at Oxford and Cambridge combined were analyzed with the configurational analysis procedure developed by Cole and Cole (1970). Briefly, the original ob-

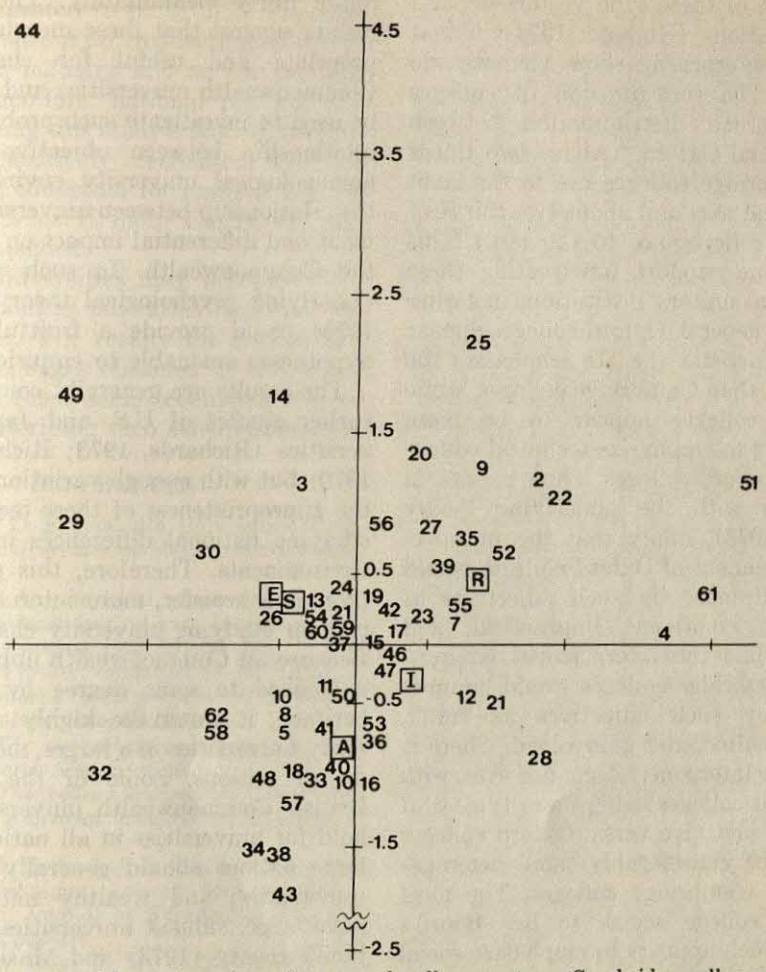


FIGURE 1. Spatial configuration of type and college vectors. *Cambridge* colleges corresponding to numbers are as follows: (1) Christ's, (2) Churchill, (3) Clare, (4) Clare Hall, (5) Corpus Christi, (6) Darwin, (7) Downing, (8) Emmanuel, (9) Fitzwilliam, (10) Girton, (11) Gonville and Caius, (12) Jesus, (13) King's, (14) Lucy Cavendish, (15) Magdalene, (16) New Hall, (17) Newnham, (18) Pembroke, (19) Peterhouse, (20) Queens', (21) St. Catherine's, (22) St. Edmund's, (23) St. John's, (24) Selwyn, (25) Sidney Sussex, (26) Trinity, (27) Trinity Hall, and (28) University. *Oxford* colleges are as follows: (29) All Souls, (30) Balliol, (31) Brasenose, (32) Christ Church, (33) Corpus Christi, (34) Exeter, (35) Hertford, (36) Jesus, (37) Keble, (38) Lady Margaret Hall, (39) Linacre, (40) Lincoln, (41) Magdalen, (42) Merton, (43) New College, (44) Nuffield, (45) Oriel, (46) Pembroke, (47) The Queen's College, (48) St. Anne's, (49) St. Anthony's, (50) St. Catherine's, (51) St. Cross, (52) St. Edmund Hall, (53) St. Hilda's, (54) St. Hugh's, (55) St. John's, (56) St. Peter's, (57) Somerville, (58) Trinity, (59) University, (60) Wadham, (61) Wolfson, and (62) Worcester.

servation vectors were located in a (reduced) space defined by principal components. This space was then projected onto the best fitting plane to provide a "picture" of the spatial configuration of the original vectors.[5] The distances between points in this picture are direct measures of the similarity between the same points (that is, similar points are close together while dissimilar points are far apart).

Figure 1 shows the resulting configuration of the vectors for Cambridge and Oxford colleges, as well as the vectors corresponding to the Holland types. The

[5] The author is grateful to Nancy Cole for carrying out the computations.

configuration of these type vectors departs from expectations (Holland, 1973) in that social and enterprising show virtually no separation. The configuration of colleges reveals a broad discrimination between Cambridge and Oxford. (About two thirds of the Cambridge colleges are to the right of the vertical axis and about two thirds of the Oxford colleges are to the left.) This provides some support for treating these universities as unitary institutions in earlier analyses. In general, Oxford colleges appear to be more artistic (i.e., to emphasize the liberal arts) than Cambridge colleges, while Cambridge colleges appear to be more realistic (i.e., to emphasize technical education) than Oxford colleges. These results, in combination with the underlying theory (Holland, 1973), imply that the interpersonal environment of Oxford colleges would be described more by such adjectives as complicated, emotional, impractical, and intuitive, while the interpersonal environment of Cambridge colleges would be more described by such adjectives as frank, stable, masculine, and uninvolved. There is considerable interpenetration, however, with some Oxford colleges being more typical of Cambridge, and vice versa. Oxford colleges appear to be considerably more heterogeneous than Cambridge colleges. The most distinctive college seems to be Oxford's Nuffield, which appears to emphasize social and enterprising fields (e.g., the social sciences and economics). Similarly, St. Cross and Wolfson at Oxford and Clare Hall at Cambridge are relatively distinctive, probably in an emphasis on science and technology. In view of such differences, an obvious question is the extent to which the configuration shown in Figure 1 corresponds to the experience of persons at Oxford and Cambridge.

Discussion

Overall, the results indicate that in the British Commonwealth (*a*) universities differ significantly on the environmental measures, (*b*) different measures of the university environment are meaningfully related to each other, and (*c*) university characteristics and characteristics of nations correlate fairly meaningfully. Therefore, the results suggest that these measures are appropriate and useful for characterizing Commonwealth universities and might now be used to investigate such problems as the relationship between objective and phenomenological university environments or the relationship between university environment and differential impact on students in the Commonwealth. In such studies, the underlying psychological theory (Holland, 1973) could provide a fruitful source of hypotheses amenable to empirical testing.

The results are generally consistent with earlier studies of U.S. and Japanese universities (Richards, 1973; Richards et al., 1970) but with enough variation to support the appropriateness of these techniques for studying national differences in university environments. Therefore, this study helps provide a broader, more international context for studying university characteristics. Because all Commonwealth universities are influenced to some degree by a common heritage, it would be highly desirable to study universities in a larger, more inclusive set of nations. Some of the results for British Commonwealth universities should hold for universities in all nations; that is, large nations should generally have large universities, and wealthy nations should have large, affluent universities. Also, Holland's theory (1973) and Maslow's (1954) notion of a hierarchy of needs would lead to the expectation that universities in wealthy nations would place relatively high emphasis on artistic fields. Other results of this study might be more specific to the British Commonwealth and, therefore, might be expected to change in a broader sample of nations. For example, developing nations would be expected to place heavy emphasis on development of agriculture and major civil engineering projects, such as roads, dams, and the like. In a broader sample of nations, this might produce a substantial negative correlation between national wealth and emphasis on education for realistic fields. Similarly, the most obvious bias in this set of Commonwealth nations is the absence of nations explicitly committed to the Communist ideology. Such nations

might be expected to place more emphasis than other nations on engineering related to the needs of industry and on the physical sciences. Therefore, national political orientation (using this study's scoring system) might be positively correlated with an emphasis on education for realistic and investigative fields if more Communist nations were included.

Finally, universities may not provide a single environment but rather multiple subenvironments. The results for Oxford and Cambridge colleges obtained in this study suggest that these same techniques can be useful for describing such subenvironments.

REFERENCES

Association of Commonwealth Universities. *Commonwealth universities yearbook 1969*. London: Author, 1969.

Astin, A. W. An empirical characterization of higher educational institutions. *Journal of Educational Psychology*, 1962, **53**, 224–235.

Astin, A. W. Further validation of the environmental assessment technique. *Journal of Educational Psychology*, 1963, **54**, 217–226.

Astin, A. W. *Who goes where to college*. Chicago: Science Research Associates, 1965.

Astin, A. W. *The college environment*. Washington, D.C.: American Council on Education, 1968.

Astin, A. W., & Holland, J. L. The environmental assessment technique: A way to measure college environments. *Journal of Educational Psychology*, 1961, **52**, 303–316.

Cole, N. S., & Cole, J. W. L. *An analysis of spatial configuration and its application to research in higher education*. (Res. Rep. 35) Iowa City, Iowa: American College Testing Program, 1970.

Holland, J. L. A psychological classification scheme for vocations and major fields. *Journal of Counseling Psychology*, 1966, **13**, 278–288.

Holland, J. L. *Making vocational choices: A theory of careers*. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Holland, J. L., Viernstein, M. C., Kuo, H., Karweit, N. C., & Blum, Z. D. *A psychological classification for occupations*. (Rep. 90) Baltimore, Md.: Johns Hopkins University, Center for Social Organization of Schools, 1970.

Lonner, W. J. The Strong Vocational Interest Blank visits German, Austrian, and Swiss psychologists. *American Psychologist*, 1968, **23**, 164–169.

Maslow, A. *Motivation and personality*. New York: Harper & Row, 1954.

Pace, C. R. *College and University Environment Scales*. Princeton, N.J.: Educational Testing Service, 1963.

Richards, J. M., Jr. A study of the "environments" of Japanese universities. *Research in Higher Education*, 1973, **1**, 87–99.

Richards, J. M., Jr., Bulkeley, E. M., & Richards, B. M. Faculty and curriculum as measures of two-year college environments. *Journal of College Student Personnel*, 1972, **13**, 433–440.

Richards, J. M., Jr., Rand, L. M., & Rand, L. P. A description of junior colleges. *Journal of Educational Psychology*, 1966, **57**, 207–214.

Richards, J. M., Jr., Rand, L. P., & Rand, L. M. Description of medical college environments. *American Educational Research Journal*, 1968, **5**, 647–658.

Richards, J. M., Jr., Seligman, R., & Jones, P. K. Faculty and curriculum as measures of college environments. *Journal of Educational Psychology*, 1970, **61**, 324–332.

Sawyer, J. Dimensions of nations: Size, wealth, and politics. *American Journal of Sociology*, 1967, **73**, 145–172.

Stern, G. G. *People in context*. New York: Wiley, 1970.

Townsend, R. *Up the organization*. New York: Fawcett, 1971.

Viernstein, M. C. *The extension of Holland's Occupational Classification to all occupations in the Dictionary of Occupational Titles*. (Rep. 100) Baltimore, Md.: The Johns Hopkins University, Center for Social Organization of Schools, 1971.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally, 1966.

(Received September 24, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 580–583

UNFAMILIAR STIMULUS TERMS IN CHILDREN'S PAIRED-ASSOCIATE LEARNING

ROBERT E. DAVIDSON[1]

University of Wisconsin—Madison

SHERRY A. B. PERRY AND PHOEBE K. BAKER

University of Oregon

Verbalization and/or imagery processes facilitate paired-associate learning. The degree of facilitation seems to depend on how well sentences or images unitize the stimulus and response terms. Unfamiliar or nonsense stimulus terms may not be able to play a role in such processes unless they are "concretized." A total of 64 third-grade pupils learned 10 paired associates with nonsense words serving as stimuli and familiar nouns serving as responses (e.g., latuk–boat). One half of the subjects looked at concretizing pictures of the paired items and heard sentences or labels only. Sentences without pictured stimuli impaired learning, while sentences with pictures facilitated learning significantly. Perhaps a sentence serves to assure the triggering of a compound image that unitizes the terms.

Verbalization and/or imagery processes facilitate paired-associate learning, often dramatically (Davidson & Adams, 1970). That assertion holds under many methodological variations, variations that have yet to be exhausted. However, despite the large numbers of studies that have been generated in this area of research (Paivio, 1971), it is not clear why the processes work so well. Indeed, it is not yet clear whether the observed facilitation follows from the operation of two independent processes (Paivio, 1971) or from a single process (Rohwer, 1972).

The research to date does seem to indicate that the magnitude of the facilitation effect varies as a function of how well the paired items are unitized, either by the experimenter or by the learner himself. Unitization under verbalization strategies usually takes the form of a simple sentence or prepositional phrase that asserts a relationship between the stimulus and response terms (Davidson & Dollinger, 1969; Rohwer, 1966). Unitization under an imagery strategy either encourages in the learner or displays for the learner spatial/dynamic/motoric relationships (Bower & Winzenz, 1970; Davidson & Adams, 1970; Rohwer, 1967; Wolff & Levin, 1972). Most of the research on verbalization and imagery processes has used as stimulus materials real words (usually nouns) that may vary on an abstract–concrete dimension. Intraverbal processes are said to be important for the learning of abstract nouns, while imagery processes are preferred in the presence of concrete nouns (Paivio, 1971). In particular, if the stimulus term is a familiar concrete noun, it can act as a "conceptual peg" to which the response term can be hooked. The conceptual-peg analysis extends to nonword consonant–vowel–consonant combinations if the experimental operations employed heighten the "concreteness" of stimuli (Philipchalk & Begg, 1971).

One aspect of the present study examines the role of "heightened" concreteness of unfamiliar stimulus terms. One half of the learners saw pictures of specially created objects that served as referents for artificial dissyllable "words" (Noble, 1952). Certainly there is nothing in cognitive theories of imagery to suggest that only semantically familiar objects can function as

[1] Requests for reprints should be sent to Robert E. Davidson, Department of Educational Psychology, School of Education, University of Wisconsin, Madison, Wisconsin 53706.

conceptual pegs. Therefore, pictorial representation of unfamiliar items was expected to boost performance considerably.

Another aspect of the study examines the role of verbalization for unfamiliar stimuli. Here, the prediction would seem to be equally clear. Verbalized "sentences" that depart from normal structures, either semantic or syntactic, do not facilitate learning (Rohwer, 1966; Rohwer & Levin, 1968). On the other hand, an artificial word serving as a stimulus term may be analogous to an abstract word, and the sentence or intraverbal assertion (i.e., the verb) might serve to attribute meaning to the unfamiliar word. In that case, sentences (as opposed to labels only) should facilitate learning.

Method

Subjects and Design

Sixty-four third-grade children, 32 boys and 32 girls, were recruited from one semirural public school. In addition, 50 third- and fourth-grade pupils were used to norm the unfamiliar, pictured stimuli. The 64 experimental subjects served in a 2 × 2 × 2 × 5 (Sex × Picture versus No Picture × Label versus Sentence × Trials) repeated measures design. The first three were between-subjects factors, and the fourth was a within-subjects comparison.

Materials

Ten dissyllable nonsense words (Noble, 1952) serving as stimulus terms were paired with 10 common object nouns. The pairs were as follows: perkil–book, bydoe–chair, latuk–boat, arkot–TV, rodig–tree, zumap–car, gokem–shoe, feemit–star, neglan–bed, and ookus–house.

For the verbal elaboration conditions, the pairs were embedded in simple declarative sentences with stimulus and response terms functioning as subject and direct object, respectively. Transitive verbs from a list of 1,000 common words were assigned at random to the paired words (e.g., "The perkil builds the book"; "The feemit keeps the star"; etc.).

For the pictorial conditions, line drawings were produced for both the stimulus and response terms. In the case of the stimulus terms, pictures of imaginary objects were created and paired at random with the nonsense labels. Pictures of the paired items were displayed side by side on 21 × 28 centimeter manila cards. The unfamiliar stimulus picture was always to the subject's left.

In order to assess the degree of unfamiliarity of the pictured imaginary objects, reproductions of the stimulus cards which had response lines added were given to pupils in third–fourth-grade combination classes. The subjects were instructed to produce associations to the unfamiliar stimulus picture (i.e., "Write down the words that come to your mind when you see it. What words does it make you think of?"). The subjects were told to "ignore" the familiar response picture which was also displayed. Asking the subjects to produce responses when both pictured objects were present was thought to approximate more closely the conditions under which the experimental pictorial groups learned the materials. The norming subjects were given 30 seconds per picture to write their responses. Two associative meaningfulness scores were calculated: Noble's (1952) m, the number of free associates in a unit of time, and Howe's (1972) N, the number of different free associates. The range of m values was 1.36–2.44 ($M = 2.07$); the range of N was 12–42 ($M = 24.2$). By way of example, for N the object labeled "latuk" produced 15 different associates with "rope" as a primary.

Procedure

The subjects participated individually. Instructions and practice with two example pairs were given to each subject prior to the experiment. The examples used only familiar object names (and pictures for the appropriate groups). The instructions pointed out specifically that the actual learning pairs would include names (or names and pictures) of "things" that the subject had never heard of (seen) before. Subjects were given successive study–test trials to a criterion of 9/10, but a minimum of 5 and a maximum of 10 trials were administered. Order of presentation of pairs was varied randomly over trials. In all conditions, the experimenter orally presented the labels or sentences, showing the pictures when appropriate for those conditions, and the subject repeated aloud what the experimenter said. The rate of presentation was 5 seconds for both the study and test portions of each trial, with no intratrial interval. At the end of the test portion of each trial, the subject was told how many correct responses he had made during that trial but not which items were correct.

Results

Numbers correct at each trial and percentages of response intrusions were the dependent measures. For items correct, sex was not a significant main effect ($F < 1$). The main effect for type of verbalization (label or sentence) was not significant ($F = 2.71$, $df = 1/56$). The pictorial factor (picture or no picture) was significant as a main effect ($F = 147.51$, $df = 1/56$, $p < .001$), and it interacted with trials ($F = 15.30$, $df = 4/224$, $p < .001$). Providing pictorial representations of the unfamiliar

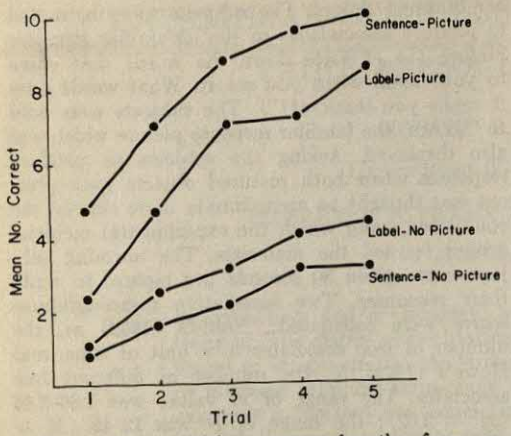


FIGURE 1. Acquisition curves for the four experimental conditions.

stimulus terms produced greater facilitation, and the facilitative effect increased differentially over trials.

The pictorial and verbalization factors interacted significantly ($F = 15.47$, $df = 1/56$, $p < .001$). The nature of the interaction may be inferred by an inspection of Figure 1; that is, sentence elaboration of pictured items produced the best performance in acquisition, while sentences in the no-picture condition produced the poorest performance.

Percentages of response intrusions, based on opportunities (total presentations minus correct responses), were analyzed following arc sine transformation. Only the main effect for type of verbalization was significant ($F = 5.73$, $df = 1/56$, $p < .05$). Percentage of response intrusions for labels and sentences was 51.5 and 40.1, respectively. Sentences act to reduce significantly the number of overt errors. A similar finding was reported by Davidson and Adams (1970).

It was of interest to analyze the relationship between learning and the associative meaningfulness measures m and N, especially in view of Howe's (1972) assertion that the use of N permits prediction of learning according to the function: Learning $= f(1/N)$. Spearman rank correlations were calculated between stimulus item rank on each meaningfulness measure and item rank in acquisition, that is, the number of times the item was given correctly in the first three trials of learning (picture groups only). Neither correlation was significant [$r_s(m) = .24$ and $r_s(1/N) = -.51$].

DISCUSSION

Certainly the most interesting effect found in this study is the significant interaction between the pictorial and verbalization conditions. The "worst" condition for learning is sentences alone, without pictures, but the "best" condition is sentences with pictures. It could be hypothesized that under a sentence–no-picture condition, the subject attempts to generate his own mediational devices. Not only are sentences provided by the experimenter not helpful to the subject, but they might also actually interfere with his attempt to form his own mediational devices. Indeed, the sentences may interfere with an out-and-out rote process. However, when a sentence is provided in the context of a picture (sentence-picture), the sentence may act to unitize the concretized terms by way of a compound image. Unitization of the pictured pairs may occur without sentences, but perhaps sentences serve to assure the triggering of the compound image that facilitates learning.

Sentences were also shown to reduce overt errors, but why sentences act to reduce overt responding is not clear. It may be the case that a conceptual-peg imagery encoding generates overconfidence in the subject which allows him to respond in the "sure" knowledge that he is correct when in fact he is not. However, imagery with concomitant sentence encoding, specifically the transitive verb, may act to limit the subject's responses to a smaller or more discriminable pool of items. That is, the verb might allow the subject to perform an editing function on an incipient response, which assures him that the response is incorrect, and he omits it. A verb constraint explanation is not entirely satisfying, however, in view of the results reported by Rohwer (1967) wherein a constraint hypothesis was essentially ruled out as an explanation for the sentence facilitation effect in paired-associate learning.

REFERENCES

Bower, G. H., & Winzenz, D. Comparison of associative learning strategies. *Psychonomic Science*, 1970, **20**, 119–120.

Davidson, R. E., & Adams, J. F. Verbal and imagery processes in children's paired-associate learning. *Journal of Experimental Child Psychology*, 1970, **9**, 429–435.

Davidson, R. E., & Dollinger, L. E. The syntactic facilitation of paired-associate learning: Deep structure variations. *Journal of Educational Psychology*, 1969, **60**, 434–438.

Howe, E. S. Number of different free associates: A general measure of associative meaningfulness. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 18–28.

Noble, C. E. An analysis of meaning. *Psychological Review*, 1952, **59**, 421–430.

Paivio, A. *Imagery and verbal processes*. New York: Holt, Rinehart & Winston, 1971.

Philipchalk, R. P., & Begg, I. Context concreteness and form class in the retention of nonsense syllables. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 499–505.

Rohwer, W. D., Jr. Constraint, syntax, and meaning in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, 1966, **5**, 541–547.

Rohwer, W. D., Jr. *Social class differences in the role of linguistic structures in paired-associate learning: Elaboration and learning proficiency*. (Final Report on Basic Research Project 5-0605) Washington, D.C.: U.S. Office of Education, 1967.

Rohwer, W. D., Jr. Decisive research: A means for answering fundamental questions about instruction. *Educational Researcher*, 1972, **1**, 5–11.

Rohwer, W. D., Jr., & Levin, J. R. Action, meaning, and stimulus selection in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, 1968, **7**, 137–141.

Wolff, P., & Levin, J. R. The role of overt activity in children's imagery production. *Child Development*, 1972, **43**, 537–547.

(Received October 2, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 584-590

VALIDITY OF STUDENT EVALUATION OF TEACHING AND THE CHARACTERISTICS OF SUCCESSFUL INSTRUCTORS[1]

ARTHUR M. SULLIVAN[2] AND GRAHAM R. SKANES

Institute for Research in Human Abilities, Memorial University of Newfoundland, Canada

Students in 10 first-year courses were asked to rate their instructors. The mean ratings for each class were correlated with the mean class mark on final, common, board-marked examinations. The mean correlation was +.39, and correlations were both positive and higher than +.32 in all but 2 of the courses. Correlations were higher for experienced full-time faculty members than for inexperienced faculty members and lowest for inexperienced part-time instructors. Academically successful and highly evaluated instructors were both "task oriented" and interest arousing. Unsuccessful but highly evaluated instructors attempted to arouse interest without being task oriented. However, electing to take subsequent courses in the subject and the level of achievement in these courses was more highly related to subject's level of achievement in the first course than to evaluation of instructor.

Although considerable data have been collected concerning the reliability of student ratings of instructors, relatively little information is available concerning the validity of student evaluations when the criterion is the amount which students learn from a course. Most studies of validity have used correlations with peer ratings or supervisor ratings as the criterion.

When final grade is used as the measure of achievement, approximately half of the studies reported show positive correlations between grades and student evaluations while the remaining half show no correlation or negative correlations between the two (Costin, Greenough, & Menges, 1971).

When a more objective measure of final achievement has been used as the criterion of achievement, for example, score on an achievement test, positive but modest correlations have usually been reported (Elliott, 1950; Morsh, Burgess, & Smith, 1956).

Using a university mathematics course taught by teaching assistants, Rodin and Rodin (1972) found, however, a correlation of −.75 between a rating of overall performance and final grade, as determined by the number of criterion problems passed, and with the effect of prior achievement removed. They concluded that students are not able to judge teaching effectiveness if the latter is measured by how much they have learned. This conclusion has been challenged by Gessner (1973) who found, in the case of sophomore medical students taking a one-semester science course, correlations of .74 and .62, respectively, between student rating of content and organization and of presentation and class performance on national normative examinations. However, no significant correlation was found between student ratings and performance on institutional examinations. The Rodin and Rodin study has also been challenged by Frey (1973), who found a positive correlation between instructor's rating and students' examination performance of .958 and .951 for two courses in mathematics.

In addition to the contradictory nature of these reports, it is difficult to generalize

[1] The research reported in this paper was supported by Canada Council Research Grant S.71-1869.

[2] Requests for reprints should be sent to Arthur M. Sullivan, Department of Junior Studies, Memorial University of Newfoundland, St. John's, Newfoundland, Canada.

from the studies which have been reported because in most cases only one course was investigated and in many cases the instructors were teaching assistants rather than full-time experienced instructors. Also, the measures of achievement used were often not those commonly used or readily understood.

At Memorial University of Newfoundland, Canada, we have been able to take advantage of circumstances which enabled us to carry out a study involving a large number of instructors, many of whom were full-time experienced faculty members, and an objective and very common measure of achievement.

This study was designed to find out if, under these circumstances, students were able to identify those instructors from whom they learned most.

A second consideration was to find out the qualities which differentiated the successful and highly evaluated instructors from those who were not successful or highly evaluated.

Method

Setting and Materials

The Junior Division of Memorial University is an administrative unit within the university which is responsible for the planning, presentation, and evaluaton of first-year university and special preparatory (foundation) courses. Foundation courses are offered in those subjects (English, mathematics, physics, chemistry, biology) in which it is important that the student have a good background before beginning a university level course but in which, because of variations in high school facilities and instruction, some incoming students have either no prior preparation or an inadequate one.

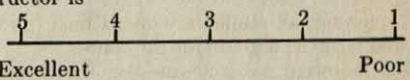
Successful teaching experience is weighed heavily in making appointments to the division and, therefore, most faculty members have considerable experience in teaching in the high school system. Some faculty members have extensive experience elsewhere but have little experience in our high school system, and a few faculty members have no previous teaching experience. In one subject (psychology), a number of graduate students are employed as part-time teaching assistants.

Although the first-year intake is in excess of 2,000 students, a sufficiently large number of faculty members have been hired to permit small classes. Students are usually taught in classes or sections with between 30 and 40 members. All subjects have several sections and those with the largest enrollment (English, psychology) have as many as 45 separate sections and 20 different instructors.

Because of the large number of sections and because of the importance of maintaining standards at the first-year level, many first-year subjects have examination committees. These committees usually include some Junior Division faculty members and some Senior Division faculty members who are involved in the teaching of the second-year courses. Early in each semester, the examination committee makes up and circulates a model of the final examination, but the final examination itself is not circulated and its specific content is not known to any except the committee members until the examination is written. The committee also sets out guidelines to be used on marking each answer. All answers are board-marked with a small group of faculty members marking one question on all papers. The marks on the final examination are then tabulated simply by adding the raw scores on each question. The mark on the final examination counts 50% of the final grade. The remaining 50% is determined by term work and is assigned by the instructor. The mark received on the final examination for those subjects in which board-marking is used is, therefore, probably as objective a measure of student achievement as it is possible to get in an applied learning situation.

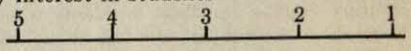
During the first semester of the 1971–1972 academic year, a student evaluation form was prepared. The form contained an evaluation section which included the following questions:

1. As an overall rating, would you say the instructor is

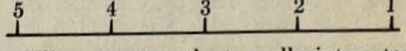


2. Specifically, how would you rate his (her):

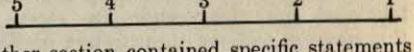
(*a*) interest in students



(*b*) ability to present material in a clear and easily understood manner



(*c*) ability to get students really interested in the subject



The other section contained specific statements which described the performance of the instructor and were intended primarily to provide feedback to him.[3]

Therefore, it was possible to obtain from each student a relatively objective measure of achievement, that is, the final examination mark; and since the rating had been done anonymously, we

[3] Copies of the evaluation form may be obtained from the senior author.

TABLE 1

CORRELATION BETWEEN INSTRUCTOR RATING AND FINAL EXAMINATION ACHIEVEMENT (DECEMBER 1971)

Subject	No. sections	r
Biology 100F	6	−.2767
Biology 1010	14	.4157
Chemistry 100F	8	.0785
Chemistry 1000	6	.5535
Mathematics 100F	8	.4772
Mathematics 1010	16	.3379
Mathematics 1150	9	.3270
Physics 1050	9	.5720
Psychology 1000	40	.4009**
Science 115A	14	.5106*
Total		+.394**

* $p < .05$.
** $p < .01$.

were able to find the mean rating for each section on the item which measured overall competence, the three items which measured specific characteristics, and the mean proportion of students who answered yes to each of the descriptive items.

Procedure

In September of 1971, approximately 2,300 first-year students registered and selected their courses. In each course, students were assigned randomly to one of the available sections. During the tenth week of the semester, the course evaluation form was completed anonymously. At the end of the 13-week semester, all students wrote a final examination and received a grade for the course.

The following subjects used a final, common examination which was board-marked and which counted 50% toward the final grade: Biology 100F, Chemistry 100F, Mathematics 100F (special preparatory courses for those students whose Grade 11 mark in the subject was less than 70% or who had not taken the subject in high school); Biology 1010, Chemistry 1000, Mathematics 1010, Physics 1050, Psychology 1000 (regular first-year courses); and Mathematics 1050, Science 115A (special first-year courses for those intending to become teachers of primary and elementary grades).

The final examination mark of each student was tabulated. For each section, the mean mark on this measure and the mean rating on Question 1 (overall competence) of the evaluation were calculated. Correlations were then obtained between mean instructor rating and mean final examination marks.

Since students had been assigned to them on a random basis, it was reasonable to assume that sections were similar in initial achievement, and in fact, remarkably little variability among sections was found in average high school marks in those subjects for which these marks were readily available.

RESULTS AND DISCUSSION

The correlations between student ratings and final examination marks are given in Table 1.

Of the 10 correlations, it can readily be seen that 9 are positive and 8 are above +.32. The average correlation is +.39. A test for significance of a number of independent tests of the same hypothesis (Winer, 1971, p. 49) revealed a significant overall positive effect ($\chi^2 = 44.13$, $df = 20$, $p < .05$). In general, then, these results indicate that there is a modest, but significant, relationship between student evaluation of instruction and student achievement.

The relatively small magnitude of the correlation coefficient in this study may be partially accounted for by the restricted range for both variables. Faculty members were, in general, evaluated highly, with little variability. The mean rating was 3.7 and the standard deviation of the ratings was .41. Similarly, the range of achievement among the sections was narrow. For most of the courses, then, the instructors formed a homogeneous group, yet although the overall significant positive correlation was low, it was a clear indication of the ability of students to identify those teachers from whom they learn most.

In subjects where the population of instructors was more heterogeneous (e.g., Science 115, in that three of the eight instructors were teaching in our university for the first time), the correlation was relatively high, indicating that in this situation students find it easier to make a valid assessment. Further support for this interpretation and additional information concerning the factors involved in the validity of student evaluations may be found in an examination of the results in Psychology 1000. This subject has the largest number of sections and the greatest heterogeneity among instructors. Of the 40 sections included in this report, 27 were taught by full-time instructors and the remaining 13 by graduate students who were hired as part-time teaching assistants. The course consists of 10 modules which are taught in a highly structured manner with clearly specified objectives and frequent evalu-

ations (Sullivan, 1969), but each instructor has considerable responsibility for the specific instructional material which is used and the method of presentation.

The relationship between evaluation and student achievement was tabulated, and correlation coefficients were calculated for the entire group and separately for the subgroups of full-time instructors and graduate students. For the group of full-time instructors, the correlation is +.528, which is significant at the .01 level. For the graduate students, the correlation is +.007 which does not differ significantly from zero. The overall correlation between rating and final examination mark for all 40 sections is +.401, which is significant at the .01 level.

The difference between the validity of ratings for part-time and full-time instructors is important and interesting. The two factors which appear to be of most importance in explaining the difference are the nature of the commitment to teaching and the amount of experience.

Full-time instructors are heavily committed to teaching and must accept major responsibility for the outcome of their instruction. A high level of student achievement is, therefore, a primary goal for them, and since they have considerable autonomy in the actual process of instruction, they are able to accept a high degree of responsibility for the attainment of this goal. Part-time teaching assistants and graduate students are not heavily committed to teaching and do not usually have much autonomy in planning the process of instruction or in carrying out their specific responsibilities. Student achievement may not be a primary goal for them. They may be more interested in alternative goals, such as getting along well with the students they teach or arousing interest in stimulating topics which are not necessarily relevant to the major objectives of the course. Student evaluation, therefore, might be based on the success that they have in attaining these other goals rather than on achievement.

This point has also been noted by Gessner (1973), who commented on the minimal level of responsibility for student achievement which characterized the teaching assistants in the Rodin and Rodin (1972) study.

The other factor, amount of experience, is also of considerable importance. When we divided our full-time instructors in psychology into those who were in their first year of full-time teaching (inexperienced) and those with one or more years of full-time teaching (experienced), we found that the correlation between evaluation and achievement was significant ($p < .01$) for the experienced ($r = .685$) but not the inexperienced ($r = .132$) instructors. This finding suggests that valid ratings are easier to obtain in the case of experienced instructors who presumably have developed a more consistent style of teaching than inexperienced instructors.

These data may help to provide an explanation of the contradictory results which have been reported in previous studies. The Rodin and Rodin (1972) results are based on part-time teaching assistants, while those of Gessner (1973) and Frey (1973) are based on full-time experienced instructors. It is likely that most of the studies which have reported no correlation or a negative correlation between student evaluation and achievement have used results from part-time or inexperienced instructors, while those which have reported positive correlations have gathered data from full-time and experienced faculty members.

On the basis of the results of this and other studies referred to above, it appears that under certain circumstances students are able to provide an accurate estimate of the amount they learn from an instructor and that student evaluations may provide a valid indication of the amount which the students have learned. Valid ratings are much more common and are easily obtained in the case of experienced and full-time instructors than in the case of inexperienced or part-time instructors.

In an attempt to find the characteristics associated with effective and highly evaluated teaching, the instructors in those subjects which have the largest number of sections (biology, mathematics, and psychology) were divided into four groups at the medians of the achievement and evalu-

ation dimensions. Of the resulting four groups, two groups, the high-achievement–high-evaluation group and the low-achievement–low-evaluation group may be described as "consonant." The other two, the high-achievement–low-evaluation group and the low-achievement–high-evaluation group may be described as "dissonant."

An examination of the mean scores for the instructors in each quadrant on the specific evaluation questions (i.e., interest in students, clarity of presentation, and ability to arouse interest) and the percentage of students answering yes to each of the descriptive items revealed the following:

1. Positive answers to items associated with attitude toward students, for example, "is friendly and understanding" and "is usually available for help," had surprisingly little relation to either favorable evaluation or a high level of student achievement.

2. Positive answers to items associated with clarity of presentation, for example, "is well prepared," uses the blackboard effectively," and "explains things clearly," were related to a favorable evaluation but were not necessarily associated with a high level of student achievement.

3. Positive answers to items associated with "task orientation," for example, "expects too much from students" and "lectures present too much material," were not related to a favorable evaluation but were associated with a high level of student achievement.

Incidentally, the items associated with a high level of achievement differed somewhat from subject to subject. In psychology and biology, success was associated with direct and continuing pressure for the student to work hard, whereas in mathematics, success tended to be associated with a supportive orientation which helped to prevent the students from becoming discouraged.

When a comparison is made for all subjects between the "consonant" instructor groups (i.e., those who are high on achievement and evaluation versus those who are low in both), it is apparent that successful and highly evaluated instructors are able to combine "task orientation" with the qualities associated with high evaluation. Thus, successful instructors are able to combine task orientation with the following: clarity of presentation, the encouragement of independent thinking, the expression of different points of view, and the ability to convey enthusiasm for the subject. Unsuccessful and poorly evaluated instructors on the other hand are obviously not able to present the material clearly, to arouse enthusiasm for the subject, or to induce the students to work hard.

When a comparison is made between the "dissonant" instructor groups, that is, those high on one dimension but not on the other, interesting findings emerge. Some examples of these are as follows.

Positive answers to questions such as "conveys enthusiasm for his subject" and "encourages the expression of different points of view" tend to be associated with a low rather than a high level of achievement. These "dissonant" instructors, then, were not able to combine arousing enthusiasm for the subject with task orientation. Some, the high-achievement–low-evaluation group, placed most emphasis on student achievement while others, the low-achievement–high-evaluation group, placed greater emphasis on arousing interest and enthusiasm.

Students may at times, often in the case of a "dissonant" instructor, have an unfavorable reaction to a successful and achievement-oriented instructor. He may be seen as a hard and demanding task master who discourages questions and comments about interesting but irrelevant material in his determination to ensure that all of the required material of the course is covered and that high standards are maintained.

On the other hand, the highly evaluated but academically less successful instructor obviously impresses the students by the efforts which he makes to arouse interest and enthusiasm and to generate discussion, but he does so at the expense of the standard of work which he requires of the students. Although it is difficult to relate these findings to those obtained from other studies, since such data are rarely reported, these results are similar to and entirely consistent with those reported by Peck and Veldman (1973).

Many will argue that it is of more

fundamental importance for the students to acquire an interest in the subject in a first course than for them to acquire information or knowledge. The argument states that students who acquire information but not interest are "turned off" from the subject while those who acquire an interest are "turned on" and, therefore, elect to do further courses in the subject and to attain a higher level of achievement in these subsequent courses.

It is, of course, difficult to provide evidence which is crucial to this argument, but we were able to gather some relevant data by examining the second-year academic performance of those students who had completed and passed Psychology 1000 in 1971–1972. Table 2 gives the percentage of students taught by each of the four groups of instructors who elected to take a second-year course in psychology and the average mark obtained in these second-year psychology courses.

The proportion of students taught by a high-achievement instructor who elect to take further courses in psychology compared with that of those taught by a highly evaluated instructor is higher although not significantly so ($F = 4.10, df = 1/18, .05 < p < .10$). The mean marks in subsequent psychology courses are more closely related to previous achievement than to instructor evaluation ($F = 9.04, df = 1/99, p < .01$).

It is interesting to note that the one group which demonstrates both the highest percentage of students continuing and the highest marks in second-year courses is that of students taught by high-achievement–low-evaluation instructors.

When the specific courses for which students registered in the second year were examined, it was found that a smaller percentage of students taught by high-achievement–low-evaluation instructors tended to register for second-year major courses in psychology than of students taught by low-achievement–high-evaluation instructors (5.3% versus 8.4%). However, even in these major courses, the marks obtained by the students of high-achievement–low-evaluation instructors were higher (61.0% versus 56.5%). Also, a higher proportion of students taught by low-achievement–high-evaluation instructors dropped psychology major courses after registration.

TABLE 2

PERFORMANCE OF STUDENTS IN SECOND-YEAR PSYCHOLOGY COURSES ASSOCIATED WITH VARIOUS TYPES OF INSTRUCTOR CHARACTERISTICS

Characteristics of instructors			Academic performance in second-year psychology courses	
Achievement	Evaluation	*n*	% students taking courses	*M* mark
high	high	8	40.7	61.5
high	low	4	43.3	66.1
low	high	5	30.8	57.9
low	low	7	26.1	55.8

It seems, then, that in introductory psychology, at any rate, if an instructor concentrates on producing a high level of achievement, his students are at least as likely to take subsequent courses and are more likely to do well in those courses than they would have if his emphasis had been on arousing interest and enthusiasm in the subject. On the other hand, the instructor who concentrates on arousing interest in the subject without at the same time taking steps to ensure a high level of achievement may be doing his students a disservice in that they may elect to major in the subject, but lacking the necessary background, they may do poorly in or fail the required courses at the second-year level.

This finding may be of considerable importance since most beginning instructors tend, and are in fact encouraged, to spend a great deal of time finding ways of making the subject interesting but do not spend much time in finding ways to encourage the students to work hard and are apprehensive about insisting on high standards of performance.

Our results certainly suggest that because students are likely to be initially interested in a subject such as psychology and because they have had no previous experiences with the subject which were discouraging, the instructor would be well advised to concentrate on student achievement and not to be overly concerned with arousing interest in

the subject. It is of crucial importance, however, that this result should not be over generalized since in a subject for which initial interest is not likely to be high and in which the student may have had previous experiences that were discouraging, for example, mathematics, the interest-arousing role of the instructor may be of much more importance.

REFERENCES

Costin, F., Greenough, W. T., & Menges, R. J. Student rating of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, 1971, **41**, 511–535.

Elliott, D. H. Characteristics and relationships of various criteria of college and university teaching. *Purdue University Studies in Higher Education*, 1950, **70**, 5–61.

Frey, P. W. Student instructional ratings and faculty performance. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, February–March 1973.

Gessner, P. K. Evaluation of instruction. *Science*, 1973, **180**, 566–569.

Morsh, J. E., Burgess, G. G., & Smith, P. N. Student achievement as a measure of instructor effectiveness. *Journal of Educational Psychology*, 1956, **47**, 79–88.

Peck, R. F., & Veldman, D. J. Personal characteristics associated with effective teaching. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, February–March 1973.

Rodin, M., & Rodin, B. Student evaluations of teachers. *Science*, 1972, **177**, 1164–1166.

Sullivan, A. M. A structured individualized approach to the teaching of introductory psychology. *Programmed Learning*, 1969, **6**, 231–242.

Winer, B. J. *Statistical principles in experimental design*. (2nd ed.). New York: McGraw-Hill, 1971.

(Received October 8, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 591-598

MULTIDIMENSIONAL SCALING OF CONCEPT LEARNING IN AN INTRODUCTORY COURSE[1]

HOWARD WAINER[2] AND KENNETH KAYE

The University of Chicago

Using a model for individual differences in multidimensional scaling (INDSCAL), 16 concepts dealing with developmental psychology were scaled before and after a course. It was found that 3 dimensions emerged, but that 1 of them was relatively unimportant for the course. The postadministration indicated that this dimension was perceived as having lesser importance. The criterion measure used was the instructor's position in the subject space; movement toward him by the students was significant from pre- to postcourse administration. Implications of this methodology for educational evaluation are discussed.

A major goal of any course of instruction is the integration of concepts into a cohesive structure. The recall of facts and the ability to define concepts are fairly easy outcomes to assess, but the extent to which students understand interrelationships among the facts and concepts is problematic. Relationships are more difficult to define; there is far less agreement among instructors and among authors as to the meaningful structure of the subject matter; and the instructor is usually ambivalent about whether his students should be acquiring *the* structure, *his* structure, or their *own* structure. One can make inferences about the body of knowledge sampled by factual or definition questions (e.g., Lord & Novick, 1968), but instructors desiring a more qualitative assessment ordinarily use essay examinations. The experience of preparing for and writing essays may be valuable to the students, but as measures of instructional effectiveness they are nonobjective, difficult to score, and relatively unreliable. (Nunnally, 1970, p. 200)

Wainer and Berg (1972) used multidimensional scaling (MDS) to determine, objectively and reliably, college students' perceived structure of interrelationships among nine short stories by De Maupassant. This study used a matrix of similarities obtained from a group of advanced French majors, employing the Shepard (1962) and Kruskal (1964) method of multidimensional scaling. A second study using the same procedure (Berg & Wainer, 1973) asked two further questions. Students judged the similarity in pairs of poems by Baudelaire, before and after several weeks of lectures and discussions. In addition, the perceived structure of students who read the poetry in the original French was compared with that of students who read the same poems in translation. The results of this study were mixed. Readers of the translations apparently perceived only one dimension dealing with ideas and imagery, while readers of the original French poems perceived a second dimension involving the rhetoric and sound patterns. With only nine poems as stimulus items, multidimensional scaling would be unlikely to establish more than two dimensions reliably. The comparison of structures before and after instruction also yielded only gross effects, and the ob-

[1] This study was made possible by the cooperation of the Developmental Psychology Laboratory, University of Washington, and by a grant from the Spencer Foundation to the University of Chicago. The authors are grateful to Gary Bond for his assistance.

[2] Requests for reprints should be sent to Howard Wainer, 5848 South University Avenue, Chicago, Illinois 60637.

servation of changes was of an exploratory, speculative nature: confirmation had to await a more delicate methodology.

The present study is an improvement in several respects: (*a*) the intervening treatment was a full course of study, including films, discussions, demonstrations, lectures, reading, and examinations; (*b*) stimulus items were single concepts, making it possible to include enough (16 items) so that several stable dimensions could emerge (Klahr, 1969); (*c*) the concepts were drawn from a longer list which formed the explicit curriculum from the students' point of view: their grades depended upon examinations consisting solely of pairs of items ("define and relate") drawn from the same list of 80 concepts; (*d*) it was possible to determine the position and movement of individual subjects in the dimensional space of the groups as a whole; and (*e*) it was possible to determine the extent to which the instructor himself perceived the same dimensions as those of the students.

The latter two improvements resulted from the use of a powerful tool: Carroll and Chang's (1970) model for individual differences in multidimensional scaling (INDSCAL), a special case of Tucker's three-mode factor analysis model (Tucker, 1972). Their model yields essentially the same spatial configuration of items as the Shepard-Kruskal method, but in addition, it uses each individual subject's judgments to weight, for him alone, the importance of each of the obtained dimensions. This is not the same as obtaining each subject's personal dimensions, but in some respects it is better. It allows us to plot individual subjects in the same space as the items, where the coordinate of a particular subject on a particular dimension is the weight he attaches to that dimension. Moreover, it calculates the proportion of each individual's variance accounted for by the structure obtained from the group as a whole.

Method

Subjects were 45 undergraduates drawn randomly from 90 who completed the introductory course in developmental psychology at the University of Washington in the summer quarter, 1972. The text (Johnson & Medinnus, 1969), films, and live demonstrations were supplementary in the instructor's mind, to the 20 lectures and the weekly discussion sections taught by a graduate assistant. Both the lectures and the discussions focused upon key concepts presented to the students in the syllabus at the first meeting, with the understanding that each of their four half-hour examinations would consist of five pairs drawn from this list. This method had been used by the instructor in three previous undergraduate courses at the University of Washington; its advantages are that students know what to expect without knowing the exact questions (they are told to expect pairs like "assimilation versus accommodation" rather than distant ones like "sex role versus formal operations"), their responses are short and pertinent to what the instructor hopes to convey, and they can be graded reliably.

Our multidimensional-scaling procedure involved only 16 of the concepts, chosen so as to vary along 3 dimensions which the instructor thought would characterize the structure of his course: (*a*) infancy–childhood–adolescence (phenomena), (*b*) nature–interaction–nurture (process), and (*c*) mental–libidinal–operant (theory). The 16 concepts together with the definitions handed to the subjects are as follows:

socialization—process by which the child learns to behave in accord with the expectations of his family, peers, and society;

formal operations—mental manipulations of objects and ideas, which together give adult thought its structural, logical properties;

basic trust—the first stage toward the development of a mature identity, in which the infant relies on the consistent care he receives from adults;

attention—focusing one's perception on a given stimulus or set of stimuli while ignoring others;

operant—a class of responses whose frequency of occurrence is determined by the way the responses have been rewarded or punished in the past;

reflex—unlearned, involuntary, specific response to stimulation of a given kind;

intention—motivation determined by the end to be achieved, which maintains an organism's behavior even in the absence of any physical stimulus or previous reward;

schema—the regularity underlying a class of responses, capable of being organized hierarchically toward the attainment of a goal or end state;

conservation—a system of mental operations, enabling older children to predict correctly that the amount of some substance will remain unchanged when it is poured or molded or spread out into a different shape;

regression—behavior that was appropriate at an earlier stage, but has been revived in later child-

hood or adulthood when it is regarded as pathological;

alienation—a feeling, in adolescents and adults, that they do not belong or are not appreciated in their families, peer groups, or society in general;

innate component—behavior which develops much more quickly and efficiently because of the genetic endowment of an individual or species, than it could have been acquired through learning or random variation;

sex role—the learned conformity to behavioral standards for girls and boys in a given society;

concept—rule relating a set of objects or ideas which share a common set of attributes and which often, but not always, share the same name in a language;

aggression—hostile, threatening behavior toward other organisms of one's own species; and

symbol—the signifier for a class of things, which (from about the age of two years) we differentiate from the objects for which the symbol stands.

All 120 possible pairs were presented in a random-order, successive-intervals format which was the same for all subjects. In addition, 8 randomly selected pairs were repeated so that we might estimate the reliability for each subject. Subjects were asked to score each pair for similarity on a scale from 1 to 15. A score of 1 was to mean that the concepts were "virtually identical" and 15 that they were "as different as could be." The subjects took from 25 to 55 minutes to complete this questionnaire, once at the beginning of the course and again at its end, 10 weeks later. The instructor and his teaching assistant also completed the questionnaire.

Results

The sample was split arbitrarily into "exploratory data" ($n = 20$) and "confirmatory data" ($n = 25$), so that hypotheses generated by inspection of one set of data could be confirmed using the uninspected data. Moreover, previous research (Berg & Wainer, 1973; Wainer & Berg, 1972) indicated that as few as 20 subjects appeared to yield stable estimates of item coordinates. Splitting the data allows us to test this.

Since eight item pairs were repeated in the questionnaire, each subject's reliability could be estimated separately. Reliabilities varied considerably, but the median was reasonably high: about .8 in both the precourse and postcourse administrations of the questionnaire; none was less than .5.

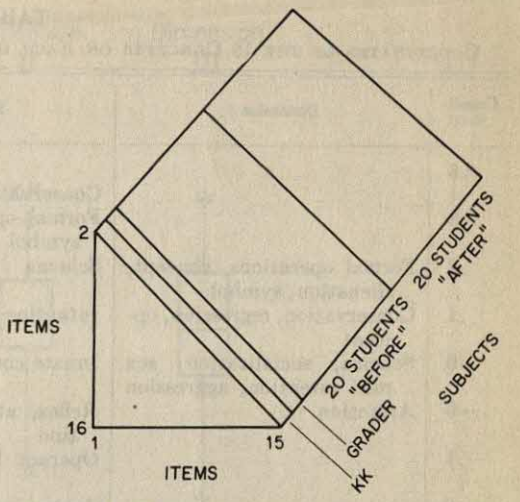


FIGURE 1. The structure of the data for the exploratory analysis. (Abbreviation: KK = Kenneth Kaye, the instructor.)

Exploratory analysis. The input to INDSCAL consisted of a three-dimensional matrix of judgments, items by items by subjects. This is most easily visualized as 42 layers of interitem distance ratings. As shown in Figure 1, one layer is that of the instructor, another consists of the judgments of the teaching assistant, 20 layers come from the students before the course, and 20 more come from the same students after the course.

A three-dimensional solution was accepted; the correlation between the data and the fit in three dimensions was .62, unusually low for INDSCAL, but about as high as could be expected given the attenuating effect of within-subjects reliability. There were 2 three-dimensional structures to be examined: first, the spatial configuration of items and, then, the positions of individual subjects within that space.[3]

The matrix of coordinates in three-space for the 16 concepts is shown in Table 1. The first dimension has *reflex, innate component,* and *basic trust* at one end, gapped (Tukey, 1971, Chapter 18x; Thissen & Wainer, 1973) from the terms relating to older children and adolescents at the other end. We interpret this dimension as "infancy–childhood–adolescence," granting

[3] These are, of course, the duals of one another.

TABLE 1

COORDINATES OF THE 16 CONCEPTS ON EACH OF THE 3 DIMENSIONS: EXPLORATORY ANALYSIS

Coordinate	Dimension I	Dimension II	Dimension III
.5			
.4		Conservation	Alienation
.3		Formal operations, concept, symbol	Aggression, reflex, regression
.2	Formal operations, concept, alienation, symbol	Schema	Innate component
.1	Conservation, regression, operant	Intention	
0	Schema, socialization, sex role, intention, aggression	Innate component	Conservation, symbol
−0	Attention	Reflex, attention, socialization	Concept, attention
−.1		Operant basic trust, sex role	Intention, formal operations, operant, schema
−.2		Aggression	Sex role
−.3	Basic trust	Regression, alienation	Basic trust
−.4			Socialization
−.5	Innate component		
−.6	Reflex		

that more advanced students would have placed *concept* and *symbol* somewhere between infancy and childhood rather than with *formal operations* and *alienation*.

The second dimension is also easy to interpret. Piagetian terms are at one end, Freudian terms at the other, and terms from learning theory fall in the middle. We call this dimension "Freud–Piaget."

The third dimension was difficult to interpret initially; perhaps the reader should refer to Table 1 and attempt his own interpretation before reading further. The instructor (Kenneth Kaye) was unable to name the dimension or to understand it with respect to his course. Table 1 was shown to two graduate students who had taken his course at the University of Chicago as well as to a colleague in the Department of Education and Human Development. For Dimensions I and II, their interpretations agreed with ours; Dimension III was labeled adjusted–maladjusted by one, prosocial–school–antisocial by another, and socialized–organismic by the third. These interpretations seem to be based primarily upon the two ends of the dimension. We shall give it the name "prosocial–antisocial." Why did the instructor have difficulty interpreting this third dimension?

In the previous section, we listed three dimensions which the instructor predicted would characterize his course. One of these corresponded exactly to Dimension I, which is, after all, an obvious dimension for a course entitled "Child Development." Dimension II, Freud–Piaget, is similar to a predicted dimension except that the instructor thought he conceived of libidinal theory as midway along a cognitive to behavioral dimension. Dimension III, too, is something like one of the predicted dimensions, nature–nurture, which suggests that the instructor should have found it easy to interpret. The predictions had been set aside for eight months, however, and were forgotten since they had been used only for reference in selecting stimulus items that would vary along some expected dimensions. We could conclude so far, then, that multidimensional-scaling structures merely reflect the experimenter's choice of stimuli. However, the dimensions become interesting when we look at how individual subjects weighted them.

Figure 2 presents "box and whisker" plots (Tukey, 1971) of the weightings of the students for each of the three dimensions.[4] For

[4] The "box and whisker" plots shown in Figure 2 are a schematic way of displaying a data distribution. The "*" indicates the extreme values of

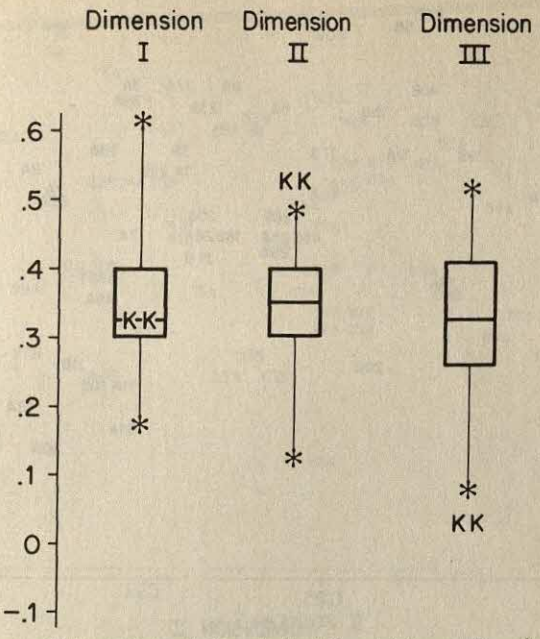


FIGURE 2. Box and whisker plots of the students' weightings of each of the obtained dimensions—an exploratory analysis. (Abbreviation: KK = Kenneth Kaye, the instructor.)

any given subject on a particular day (before or after the course), the three weightings tell us his position relative to the structure of the whole class. This gives us a way of comparing individual subjects with the instructor and testing for before–after movement through this common, conceptual reference space.

The weighting computed by INDSCAL for the instructor on Dimension I was .34, almost exactly the median of the class. On Dimension II, his weighting was .51, higher than any of the students. Although prior to the course he had not described the theoretical dimension as having Freud at the opposite end from Piaget, he used this dimension most heavily in making the paired judgments. On Dimension III, his weighting was −.03, lower than any of the students. He did not think of his course as presenting a prosocial–antisocial perspective; he did not use it in rating the concepts for similarity; and he could not even interpret Dimension III when shown Table 1.

Since Dimensions II and III were the ones on which the instructor and his student differed, we plotted their weights in Dimension II against those in Dimension III (Figure 3). In this figure, the subject identification numbers are shown along with the letter B or A for before or after. Evidently, 9 of the 20 moved toward the instructor on *both* dimensions, and another 10 moved toward him on *either* the Freud–Piaget dimension (giving it greater weight after the course) or the prosocial–antisocial (giving it less weight). More technically, the "before" centroid is significantly further from KK (Kenneth Kaye, the instructor) than the "after" centroid ($p < .05$).

Note that the graduate assistant perceived the concepts in a manner similar to that of the instructor, but still well within the range of the students. The fact that she graded the examinations may help to explain our failure to get a significant correlation between final grade and movement toward the instructor.

Confirmatory analysis. The data from 25

the distribution, read off the stem. The lines on the ends of the box represent the 25th and 75th percentiles of the distribution; thus the box encloses the middle 50% of the data. The middle line is the median. This method of data display is quite useful in that it gives the "flying eye" a picture of overall level and spread simultaneously, and this allows easy comparison across groups.

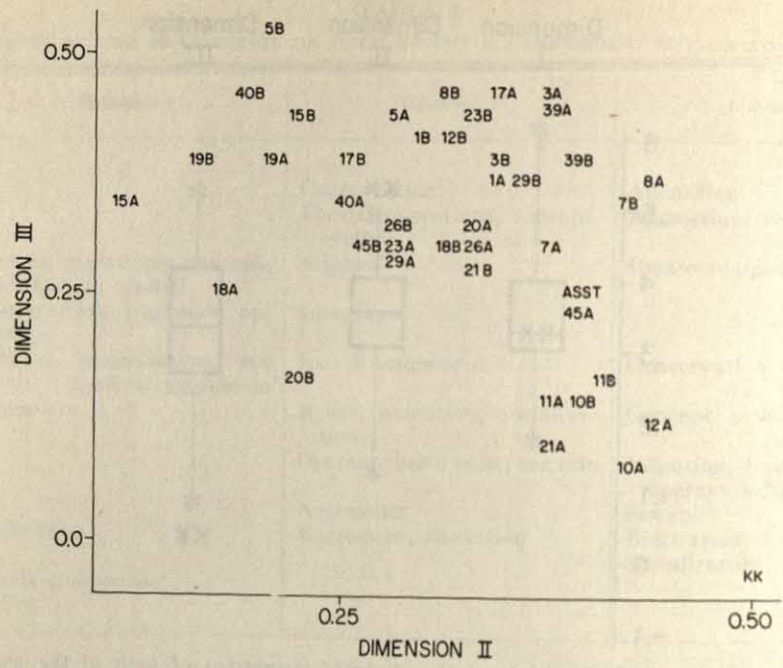


FIGURE 3. The weightings of the exploratory subjects on Dimension II and Dimension III before (B) and after (A) a course in developmental psychology. (Abbreviation: KK = Kenneth Kaye, the instructor.)

additional subjects, set aside during the first phase of analysis, were now used as input to INDSCAL. The three-dimensional solution had about as good a fit as before [r(data, fit) = .61]. The matrix of item coordinates was virtually identical to Table 1. The correlations between exploratory and confirmatory samples, for the rank order of items along each dimension, were .97, .95, and .92 for Dimensions I, II, and III, respectively. Although our interpretation and labeling of these dimensions remains subjective, the great similarity between replications of the analysis indicates that sample sizes of about 20 are sufficient to give highly reliable estimates of item coordinates and of dimensional structure in this application of multidimensional scaling.

Examination of the projections of subject scores in the Dimension II–III plane indicates that the instructor's position is, once again, rather solitary. As before, movement toward him by the students is significant ($p < .05$). Sadly, the degree of movement appears to be unrelated to grade ($r = .18$), number of previous psychology courses, or sex. We are uncertain if this reflects badly on multidimensional scaling or on the method of course evaluation which determined grades (see Figure 4).

DISCUSSION

In this paper, we have utilized an individual-difference model of multidimensional scaling to get estimates of the effect of instruction upon movement through a conceptual space. The results attest to the power of Carroll and Chang's (1970) INDSCAL model. The remarkable difference between the instructor and his students in their use of the three major dimensions, which held in replication with the confirmatory sample, was unexpected but believable. The change in perception of the class as a whole (as measured by the "before" and "after" centroids of weights), shifting significantly toward the instructors's own view of his material, was a modest but very important effect.

A subsequent study in which the instructor introduces relevant dimensions as part of the explicit content of his course should achieve more drastic changes in the students' conceptual structures. If such a study

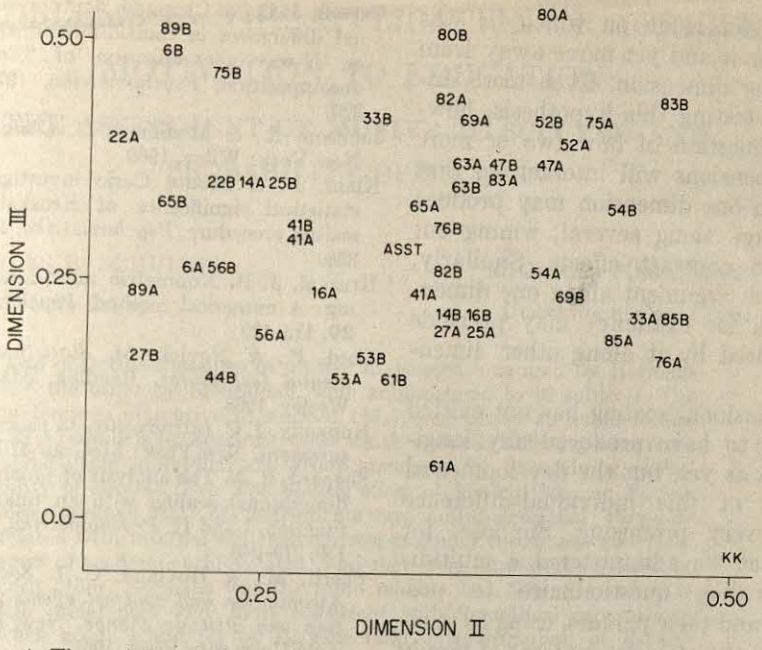


FIGURE 4. The weights of the confirmatory subjects on Dimension II and Dimension III before (B) and after (A) a course in developmental psychology. (Abbreviation: KK = Kenneth Kaye, the instructor.)

found less overlap between the "before" and "after" data points, it might be good to construct a separate multidimensional-scaling space from each of the two sets of data. In the present experiment, however, the overlap indicates that the first three dimensions used remained the same at the end of the course. Carroll and Chang's procedure enables us to trace a shift in the relative importance attached to those dimensions by the students. Infancy–childhood–adolescence was an obvious continuum to keep in mind when judging concept pairs. On the less obvious second and third dimensions, the course apparently influenced the students to regard the theoretical loading of a concept as more important, and the prosocial–antisocial loading as less important than they had done initially.

Another goal for future research in this area is to discover or produce a relationship between a student's movement toward the instructor and the subjective grade he receives. The implications of such a relationship are obviously complex, and it is not clear that the best students are those who assimilate their professor's biases. It is still too early to advocate multidimensional scaling for assessing students, but it does yield useful insights for the instructor.

In addition to refining the present application of multidimensional scaling, we hope to explore others. Here we have measured individual perceptions of, and changes in, the underlying structure of a course; we have used the instructor's position in the same dimensional space as a criterion measure for individual subjects. A somewhat different application would be in the study of attitude change. In the research of Sherif and Hovland (1961), the extent to which a person is influenced by an argument is curvilinearly related to the initial distance between his own attitude and the presented one. The individual-difference multidimensional-scaling procedure makes it possible to view each person within an attitude space, including the person attempting to influence the others. Thus, the assimilation and contrast effect can be seen multidimensionally. A person whose attitude is near the target on one dimension might be far from it on another; theory predicts he should be positively influenced by the argu-

ment in the dimension on which he was relatively near it and yet move away from it on the other dimension. Even more important than testing this hypothesis, however, is the question of how two or more cognitive dimensions will interact, so that persuasion on one dimension may produce positive change along several, wiping out the predicted contrast effects. Similarly, aversion to an argument along one dimension (racism, for example) may preclude being influenced by it along other dimensions.

Multidimensional scaling has not existed long enough to have produced any longitudinal data as yet, but the developmental applications of this individual-difference model are very promising. Suppose, for example, that we administered a multidimensional-scaling questionnaire to preadolescents and their parents, using as stimulus items, pictures or descriptions of family roles. The same subjects could be tested 5 and 10 years later; children should change more than adults, in the direction of their own parents more than in a chance direction, toward the same-sexed parent more than the other, and so forth.

REFERENCES

Berg, W., & Wainer, H. Baudelaire: What's lost in translation? An application of multidimensional scaling. Chicago: University of Chicago, Department of Psychology, 1973. (Mimeograph)

Carroll, J. D., & Chang, J. J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 1970, **35**, 283–320.

Johnson, R., & Medinnus, G. *Child Psychology*. New York: Wiley, 1969.

Klahr, D. A Monte Carlo investigation of the statistical significance of Kruskal's nonmetric scaling procedure. *Psychometrika*, 1969, **34**, 319–330.

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 1964, **29**, 115–129.

Lord, F., & Novick, M. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Nunnally, J. C. *Introduction to psychological measurement*. New York: McGraw-Hill, 1970.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function, I and II. *Psychometrika*, 1962, **27**, 125–139, 219–246.

Sherif, M., & Hovland, C. I. *Social judgment: Assimilation and contrast effects in communication and attitude change*. New Haven, Conn.: Yale University Press, 1961.

Thissen, D., & Wainer, H. GAPPER: An IBM 360 program for finding gaps in univariate data strings. *Behavioral Science*, 1973, **18**, 77–78.

Tucker, L. R. Relations between multidimensional scaling and three mode factor analysis. *Psychometrika*, 1972, **37**, 3–27.

Tukey, J. W. *Exploratory data analysis*. (preliminary ed.) Reading, Mass.: Addison-Wesley, 1971.

Wainer, H., & Berg, W. The dimensions of de Maupassant: A multidimensional analysis of students' perception of literature. *American Educational Research Journal*, 1972, **9**, 485–492.

(Received October 8, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 599–608

SOME PROBLEMS IN THE APPLICATION OF ACHIEVEMENT MOTIVATION TO EDUCATION: THE ASSESSMENT OF MOTIVE TO SUCCEED AND PROBABILITY OF SUCCESS[1]

CHARLES B. SCHULTZ[2]
Trinity College

MICHAEL POMERANTZ
University of Connecticut

Two objective measures of motive to succeed, one used by Hermans and the other by Mehrabian, were administered to 93 subjects. The preferences of motive to succeed (M_S) and motive to avoid failure (M_{AF}) subjects for academic tasks which varied in difficulty were compared using individual and group standards for determining probability of success (P_S). The biasing effect of M_S and M_{AF} on P_S was also examined. Both M_S instruments were multidimensional; they correlated with internal achievement responsibility for success and measures of academic achievement; and they were significantly related to each other. Hermans was the more reliable test. The M_S and M_{AF} risk preferences were most consistent with theoretical expectations when group standards of P_S were used. Overestimation of P_S was directly related to M_S. Both M_S and M_{AF} subjects overestimated P_S more on difficult tasks than on easy tasks.

Achievement motivation as conceived by Atkinson (1964, 1966) results from the interaction of three primary components, one situational and two personality variables. Individuals may be characterized by the relatively stable traits of the motive to succeed (M_S) and the motive to avoid failure (M_{AF}). For some persons, one or the other trait is dominant. Thus, a person may be basically success oriented ($M_S > M_{AF}$) or failure threatened ($M_{AF} > M_S$). Both personality types face tasks which vary in difficulty and for which they make subjective judgments of their probability of success (P_S). In addition to the primary constructs of M_S, M_{AF}, and P_S, three derivatives of P_S are assumed to account for achievement motivation. These include the incentive value of success (I_S or $1 - P_S$), the probability of failure (P_F or $1 - P_S$), and the incentive value of failure (I_F or $-P_S$).

The crux of achievement motivation theory is the notion that individuals who are highly success oriented relative to their fear of failure prefer tasks of moderate P_S, while persons with the opposite tendencies prefer either extremely easy or extremely difficult tasks. Application of this potentially useful theory to classroom instruction has been hampered by at least two assessment problems, the measurement of M_S and of P_S. These problems are complicated by the fact that risk-taking behavior across tasks is not highly correlated (Weinstein, 1969); therefore, findings using games and other tasks may not be generalizable to academic work.

Measurement of M_S, a primary problem facing educational application of the theory (Atkinson, 1965; Maehr & Sjogen, 1971), is typically determined by projective techniques such as the Thematic Apperception

[1] The authors gratefully acknowledge the assistance of John Battle and Ken Padach in the conduct and analyses of this research. They are also indebted to Robert K. Gable of the University of Connecticut for his assistance and advice regarding the factor analyses and to Charles Clock, Edward Bierman, and Francis Whittle of the West Hartford School District for their cooperation. This research was supported, in part, by a grant from Trinity College, Hartford, Connecticut.

[2] Requests for reprints should be sent to Charles B. Schultz, Department of Education, Trinity College, Hartford, Connecticut 06106.

Test and the French Test of Insight. Objective versions of tests of M_S are needed not only to increase convenience in administration and scoring but to improve upon the internal consistency and the test–retest reliabilities of the projective measures which have been low (Clarke, 1973; Weinstein, 1969). Weinstein's (1969) examination of self-report instruments found few relationships among the tests, little relationship between objective and projective instruments, and an inability of virtually all measures to predict risk-taking behavior of M_S and M_{AF} subjects on a variety of tasks. The present study is an attempt to extend Weinstein's (1969) research by assessing the reliability (internal consistency) of two more recent objective measures of M_S by Hermans (1970) and Mehrabian (1969) and by assessing their power to predict the risk-taking behavior of M_S and M_{AF} subjects on tasks that are characteristic of schoolwork. The Mehrabian instrument, in particular, was selected because of successes reported with it (Farley & Mealiea, 1973; Weiner, Johnson, & Mehrabian, 1968; Weiner & Potepan, 1970; Wolk & DuCette, 1973).

Assessment of P_S has not received the systematic attention given to the measurement of M_S, yet it is just as important, particularly because other factors in the theory are derived from it. There are at least two problems associated with the determination of P_S. The first deals with the manner in which P_S is estimated. Atkinson (1964) conceived of P_S as an individual, subjective judgment. In actuality, combinations of the various estimates of P_S have been used in research on achievement motivation. Isaacson (1964), for example, used a ratio of average grades in a course to average intelligence of the students, implying an objective, group estimate of P_S which is, in effect, the opposite of Atkinson's (1964) suggestion. A similar approach was used by deCharms and Carpenter (1968), who defined P_S by the difficulty level of arithmetic and spelling items. Frequently, P_S is induced by informing subjects of the number of others like themselves who have been successful on a task (Weiner, 1972); this also implies a group, objective standard. Weinstein (1969) used what amounted to both individual-objective and individual-subjective estimates. The present study is an attempt to determine which, if any, combination of task-specific estimates of P_S approximates the theoretical expectations derived from Atkinson's theory.

A second problem in the assessment of P_S is the tendency to bias estimates of P_S which arise from the lack of independence between M_S and P_S. Atkinson (1964) notes that persons who are high in M_S are predisposed to overestimate P_S (i.e., perceive their chances of success as greater than objective difficulty level would suggest) as compared to those who are low in M_S or who are anxious about failure. Thus, a general biasing across all levels of P_S is implied. Feather (1965) advanced two hypotheses which suggest a more complex relationship. The first is that overestimation of P_S by M_S subjects and underestimation of P_S by M_{AF} subjects will be greatest when the tasks are moderately difficult, that is, when $P_S = .50$. Feather's (1965) contention is that since M_S persons have been most successful at that difficulty level and M_{AF} persons have been relatively unsuccessful at that same level, the two groups approach moderately difficult tasks, in particular, with different expectations.

The second hypothesis is based on the assumptions that (*a*) overestimation of P_S is a function of the subjective attractiveness of success ($M_S \times I_S$) and (*b*) underestimation is the result of the subjective repulsiveness of failure ($M_{AF} \times I_F$). For persons for whom $M_S > M_{AF}$, overestimation of P_S is greatest when I_S is high and, therefore, P_S is low. When $M_{AF} > M_S$, underestimation of P_S is greatest when I_F is high and, therefore, P_S is high. Accordingly, the alternate hypothesis holds that the underestimation of P_S by M_{AF} subjects is greatest on easy tasks and that there is little, if any, underestimation on difficult tasks, while the overestimation of P_S by M_S subjects follows the opposite trend. The design of the present study provides an opportunity to examine both of Feather's hypotheses and the general positive relationship between M_S and P_S suggested by Atkinson (1964).

In summary, the purposes of the present study are to assess the reliability and the validity of two objective measures of M_S, to compare several ways of estimating P_S, and

to examine the biasing effect of M_S on P_S. Since an overriding concern of this study is the application of achievement motivation to education, academic tasks were used with a junior high school population.

Method

Subjects

The subjects were 93 male ninth graders who were randomly drawn from two suburban junior high schools. For the analyses of variance, the top quartile ($n = 23$) of each M_S scale (or resultant measure) defined the $M_S > M_{AF}$ group for that scale (hereafter refered to as M_S subjects) and the bottom quartile defined the $M_{AF} > M_S$ group (hereafter refered to as M_{AF} subjects). The other analyses were based on all subjects.

Instruments and Apparatus

Achievement measures were obtained as a part of the regular school testing program two months before the present study was conducted. Some of the personality measures were administered in their original form, while others were modified to meet the reading level and experience of ninth graders as determined by pilot work with the tests. Instruments to assess P_S and task attractiveness or preference (T_A) were constructed expressly for the purposes of the present study.

Measures of M_s

Two measures of M_S were used. The first was Mehrabian's (1969) short form of the Resultant Achievement Motivation Test. The male scale was altered as follows: four items were modified to make them less complex and, therefore, easier for ninth graders to read, and four items were dropped because of inappropriate content for the present population. Thus, the final version of the male scale consisted of 22 items. Hermans' (1970) 29-item Prestatie Motivatie Test had minor changes in verb tense and vocabulary.

Measure of M_{AF}. The Debilitating Anxiety subscale of the Achievement Anxiety Scale by Alpert and Haber (1960) served as a measure of M_{AF}. Seven of the 10 items of this subscale were changed. These modifications were slight and were limited to words which were difficult for junior high subjects as determined by pilot work with the test.

Measures of P_S and T_A. The Comprehensive Test of Basic Skills was administered two months before the present test battery. Based on that administration, 21 items were selected which varied in difficulty. Twelve items tested English skills and 9 tested mathematics. The 21 items were arranged into three categories of item difficulty based on the scores of ninth graders at the two schools. Seven were in the 20%–30% right range; 7 in the 45%–55% right range; and 7 in the 70%–80% right range. The items were arranged in sets of 3; each set tested similar skills (e.g., mathematics work problems) and contained 1 item from each level of difficulty. The set was enlarged to 5 items with the addition of an extremely easy problem (e.g., 9,605 − 4,847) and a problem that would be difficult for college students (e.g., if $dy/dx = 4(x^2 + 2)^3$, what does $(d^2y/(dx_2) = ?)$. This procedure resulted in a total of 35 items, 20 in English and 15 in mathematics, which were broken down into six sets of 5 items. Each set contained similar items with 1 item representing each of the five levels of difficulty: 0–10, 20–30, 45–55, 70–80, 90–100. The items comprising each set were printed on a separate sheet of paper in a random order. The test booklet contained four pages of English items followed by three pages of mathematics items. The booklet was presented to the subjects twice, on the first occasion to measure T_A and on the second to measure P_S.

Five-point scales were used for both P_S and T_A measures. The P_S scale reflected the five categories of the objective difficulty of the items. It ranged as follows: (*a*) I would definitely get it right (90%–100% sure of being correct); (*b*) I would probably get it right (70%–80% sure); (*c*) I could get it right or wrong (45%–55% sure); (*d*) I would probably get it wrong (20%–30% sure); and (*e*) I would definitely get it wrong (0%–10% sure). The T_A scale was as follows: (*a*) would definitely *not* want to do problem, (*b*) would not like to do problem, (*c*) don't care about doing problem, (*d*) would like to do problem, and (*e*) would definitely *want* to do problem. Thus, the T_A scale provides for both approach and avoidance tendencies.

Estimates of P_s

Three task-specific estimates of P_S were used as follows: objective group, subjective group, and subjective individual. The objective-group estimate was based on the percentage of students in the two schools who got the items right on the Comprehensive Test of Basic Skills. A priori judgments were made on the extremely easy or extremely difficult items as described above. Subjective-group estimates were the subjects' average ratings of P_S for each item. Subjective-individual estimates were simply the P_S ratings given by the subject for each item.

Procedure

The seven tests were administered in 2 one-hour sessions. During the first session, Mehrabian's (1969) Resultant Achievement Motivation Test, both Facilitating and Debilitating subscales of the Achievement Anxiety Scale (Alpert & Haber,

[3] A listing of the Prestatie Motivatie Test items according to their subscale membership and with their factor loadings and classifications according to Hermans' (1970) a priori analyses is available from the first author upon request.

1960), and a 46-item school attitude and behavior questionnaire were administered, in that order. Two weeks later, the following tests were given: the measure of T_A, the Prestatie Motivatie Test (Hermans, 1970), the Intellectual Achievement Responsibility Questionnaire (Crandall, Katkovsky, & Crandall, 1965), and the measure of P_S.

The tests were administered to groups of approximately 25. Instructions to P_S and T_A sections cautioned the subject not to work the problems but to estimate the attractiveness and later the difficulty of each item. The P_S instructions alerted the subject to the range of difficulty included in the items. The T_A instructions asked the subject to imagine that he was selecting items for a test which no one else would see but which would show him how good he was in English and mathematics. Thus, an attempt was made to establish an evaluative situation which minimized extrinsic factors such as grades.

All instruments except the measures of P_S and T_A and the school attitude and behavior questionnaire were administered via 2 × 2 inch slides. Subjects read each item projected on the screen as they listened to a tape recorded reading of the item by the experimenter. The timing of the items was determined by pilot work at a different junior high school and was controlled by an audiosynchronizer device.

Results

Assessment of M_S

The internal consistency estimate of reliability (α coefficient) for the Resultant Achievement Motivation Test (Mehrabian, 1969) was .55 and .91 for Hermans' (1970) Prestatie Motivatie Test, implying considerably more consistency for the latter measure.

Both tests were factor analyzed and the reliabilities of subscales derived from those analyses were determined. The analysis of the Resultant Achievement Motivation Test yielded 10 factors with eigenvalues greater than 1.00. These factors accounted for 70% of the total variance, while the first factor alone accounted for 12%. A varimax rotation of the 10 factors yielded 3 factors, each containing three items with factor loadings of .40 or higher, 5 factors containing two items, and 2 specific factors containing only one item. One item did not load on any factor according to the .40 criterion. The three-item factors were identified as subscales with rubrics similar to those used by Mehrabian (1968). The first (α coefficient = .57) reflected independence in two of the items (preference for working alone and for individual sports or games). The status of the third item in this subscale (preference for a book over a movie) was somewhat ambiguous. The second subscale (α coefficient = .53) was identified as pride in accomplishment due to preferences for doing things well and for love of winning. An item which tapped future orientation was also included. The third subscale (α coefficient = .52) reflected choice of achievement activities by preferences for games few people know, for difficult, thought games, and for being in business for oneself.

The Prestatie Motivatie Test (Hermans, 1970) was analyzed into nine factors with eigenvalues greater than unity; these accounted for 70% of the total variance. The first factor accounted for 34% of the variance. Five subscales that had more than three items with factor loadings of at least .40 were obtained from the varimax rotation of the nine factors. The match between Hermans' (1970) a priori judgments of the items and the subscale membership obtained from the factor analysis of the items is far from perfect. Persistence items appear in four of the five subscales. Some subscales (e.g., Subscales 2 and 5) contain no duplication of item categories and almost defy classification, while others (e.g., Subscales 3 and 4) reveal considerable overlap between Hermans' classification scheme and the factor analysis results. The first subscale contains all of the task-tension items and all but one of the aspiration-level items.

In summary, the Resultant Achievement Motivation Test (Mehrabian, 1969) is complex and multidimensional. It was analyzed into three subscales which reflect recognizable aspects of achievement motivation. However, the internal consistency of the test as a whole was relatively low, as were the reliabilities of the subscales. Moreover, 13 of the 22 items did not load on factors yielding subscales which attained even that modest level of reliability. The Prestatie Motivatie Test (Hermans, 1970) has high internal consistency, and at least three subscales are reliable. Twenty-two of the 29 items are included in the five subscales. However, there is an apparent lack of consistency between Hermans' original classifications and some of the subscales determined by the present analysis.

Intercorrelations among personality, ability, and achievement factors are listed in Table 1. The Resultant Achievement Motivation Test – Debilitating Anxiety subscale was included as an index of resultant motivation because of Weiner and Potepan's (1970) finding of only a slight relationship between the Resultant Achievement Motivation Test and test anxiety ($r = -.13$), and their subsequent argument that the Resultant Achievement Motivation Test is in actuality a measure of M_S alone. Hermans' (1970) Prestatie Motivatie Test, of course, was treated only as a measure of M_S.

The correlations on Table 1 permit several conclusions. First, if the Resultant Achievement Motivation Test is a resultant measure as Mehrabian (1968, 1969) intended it, it should correlate negatively with test anxiety; if it is a measure of M_S alone, no correlation is expected. The significant but modest correlation of $-.24$ is somewhat equivocal in this respect, and it is not substantially different from earlier correlations ($r = -.16$ and $-.26$) obtained by Mehrabian (1968, 1969). In view of this, and particularly because there are, for the most part, only trivial differences in the correlations between the Resultant Achievement Motivation Test and the Resultant Achievement Motivation Test – Debilitating Anxiety subscale with other factors, in subsequent analyses and discussions, the Resultant Achievement Motivation Test will be treated as a resultant measure. Second, all measures of M_S correlate with total internal achievement responsibility and, in particular, with internal achievement responsibility for success. No substantial correlations were obtained with internal achievement responsibility for failure. Third, all M_S indices correlate strongly with achievement, particularly when measured by grade point averages.

The Resultant Achievement Motivation Test correlates with the Prestatie Motivatie Test – Debilitating Anxiety subscale ($r = .40$, $p < .001$) somewhat more strongly than Mehrabian (1969) reports its relationship with the Thematic Apperception Test – Test Anxiety Questionnaire ($r = .29$, $p < .01$), a more traditional measure of resultant motivation. Both in spite of this and the similarities between the two M_S measures evidenced in Table 1, the Prestatie Motivatie Test can be distinguished by stronger correlations with measures of locus of control (Intellectual Achievement Responsibility Questionnaire, Total), approaching the relationship of $r = .64$ reported by Mehrabian (1968). In addition, the Prestatie Motivatie Test correlates only slightly with IQ and standard achievement tests (Comprehensive Test of Basic Skills). The first relationship is expected (Weiner, 1970, p. 219), and the second is not surprising given the strong correlation between IQ and the Comprehensive Test of Basic Skills ($r = .77$, $p < .001$). However, there is a pronounced relationship between the Prestatie Motivatie Test and grade point average, an achievement measure which may be more sensitive to motivational influences than standardized tests.

TABLE 1

CORRELATIONS OF M_S AND $M_S - M_{AF}$ WITH PERSONALITY, ACHIEVEMENT, AND ABILITY TESTS FOR MALES

Measure	Measures of M_S and $M_S - M_{AF}$			
	RAM	RAM-DAS	PMT	PMT-DAS
DAS	−.24	−.79[a]	−.28	−.80[a]
IAR-S	.44***	.47***	.70***	.62***
IAR-F	−.02	.00	.24	.16
IAR-T	.29**	.32**	.62***	.52***
IQ	.29**	.38***	.20	.31*
CTBS	.41***	.52***	.27	.42***
GPA	.48***	.51***	.56***	.57***

Note. Abbreviations: M_S = motive to succeed; M_{AF} = motive to avoid failure; DAS = Debilitating Anxiety subscale; IAR-S = Intellectual Achievement Responsibility, Success; IAR-F = Intellectual Achievement Responsibility, Failure; IAR-T = Intellectual Achievement Responsibility, Total; CTBS = Comprehensive Test of Basic Skills; and GPA = grade point average for the semester data were collected.

[a] Correlation of a composite measure with one component.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

Assessment of P_S

Analyses were conducted to determine which estimate of P_S most closely approximates theoretical expectations and to deter-

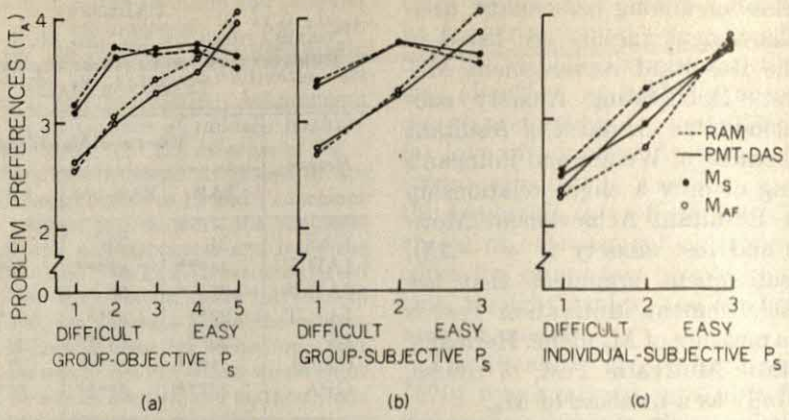


FIGURE 1. Preferences (T_A) of motive to succeed (M_S) and motive to avoid failure (M_{AF}) subjects for problems with different levels of probability of success (P_S). (In Figure 1a, the P_S levels are as follows: 1 = 0–10, 2 = 20–30, 3 = 45–55, 4 = 70–80, and 5 = 90–100; in Figures 1b and 1c, the levels are as follows: 1 = 0–10 and 20–30, 2 = 45–55, and 3 = 70–80 and 90–100. Abbreviations: RAM = Resultant Achievement Motivation Test and PMT-DAS = Prestatie Motivatie Test – Debilitating Anxiety subscale.)

mine the relationship between P_S and M_S. In regard to the first of these questions, a repeated measures analysis of variance of T_A scores was used to test the hypothesis that M_S and M_{AF} differences were greatest when $P_S = .50$. The results of these analyses are displayed in Figure 1 for each P_S estimate: group objective, group subjective, and individual subjective. Five levels of P_S (0–10, 20–30, 45–55, 70–80, 90–100) were used for the group objective. For the remaining estimates, the extreme categories were collapsed, resulting in three P_S levels (difficult = .05 and .25; moderately difficult = .50; and easy = .75 and .95).

The 2 × 5 analysis of variance of T_A scores using group-objective estimates of P_S (Figure 1a) yielded a significant main effect for P_S ($F = 16.40$, $df = 4/176$, $p < .001$) on the Prestatie Motivatie Test – Debilitating Anxiety subscale and on the Resultant Achievement Motivation Test ($F = 17.22$, $df = 4/176$, $p < .001$). The same analysis yielded a significant Personality Type × P_S interaction for the Prestatie Motivatie Test – Debilitating Anxiety subscale ($F = 4.51$, $df = 4/176$, $p < .01$) and the Resultant Achievement Motivation Test ($F = 8.10$, $df = 4/176$, $p < .001$). Although the Personality Type × P_S interactions are significant on both M_S measures, the curves clearly depart from the theoretical model in several important respects. According to pairwise comparisons of the means, the difference between M_S and M_{AF} subjects is greatest at the extreme and particularly at the difficult P_S levels, although even there the differences are less than reliable. In addition, M_{AF} risk preferences drop as P_S decreases to moderate levels as expected (according to Newman-Keuls stepwise comparison procedures with $p = .05$); however, instead of rising as P_S continues to drop, M_{AF} risk preferences also slide from $P_S = .50$ to $P_S = .00$–$.10$ ($p < .05$). The risk preferences of M_S subjects could be described as an extremely shallow, inverted U shaped function. The middle three P_S levels are higher than the hardest level ($p < .05$) and the easiest ($p > .05$) one, at least with the Resultant Achievement Motivation Test.

Similar analyses were made on the group-subjective estimates of P_S (Figure 1b). As noted above, three levels of P_S were used in this and the following analyses. Significant main effects were obtained for the P_S factor ($F = 18.02$, $df = 2/88$, $p < .001$, for the Resultant Achievement Motivation Test; $F = 12.30$, $df = 2/88$, $p < .001$, for the Prestatie Motivatie Test – Debilitating Anxiety subscale). Again, there were significant Personality Type × P_S interactions with both M_S instruments ($F = 11.84$, $df = 2/88$, $p < .001$, for the Resultant Achievement Motivation Test; $F = 5.02$, $df = 2/88$, $p < .01$, for the Prestatie Motivatie Test-

Debilitating Anxiety subscale). The same trends occurred here as noted earlier with the group-objective estimates. The only difference between personality types was at the difficult P_S level on the Resultant Achievement Motivation Test ($t = 1.70$, $df = 88$, $p < .05$). According to the Newman-Keuls procedure, M_S subjects do not differ across P_S levels; however, there is a significant ($p = .05$) increase in M_{AF} risk preference at each P_S level.

The analysis using the individual-subjective estimate of P_S (Figure 1c) yielded significant main effects for P_S on the Resultant Achievement Motivation Test ($F = 23.00$, $df = 2/88$, $p < .001$) and on the Prestatie Motivatie Test – Debilitating Anxiety subscale ($F = 18.72$, $df = 2/88$, $p < .001$). No interaction effects were obtained.

The curves in Figure 1 and these analyses can be summarized as follows:

1. Measurement of $M_S - M_{AF}$ with the Resultant Achievement Motivation Test and the Prestatie Motivatie Test – Debilitating Anxiety subscale produces nearly identical results.
2. On both group estimates of P_S, there is a direct relationship between M_{AF} and P_S, while M_S does not differ across P_S or it is depressed at the lower P_S limits.
3. There is a general, but nonsignificant, tendency for the M_S and M_{AF} curves to cross at the extremely easy P_S level.
4. When individual-subjective estimates of P_S are used, the T_A scores of both M_S and M_{AF} subjects are directly related to P_S, a finding which departs most sharply from the theoretical model.

The influence of M_S on subjective P_S judgments was assessed in two ways. The P_S ratings were summed across all 35 problems and were correlated with the measures of M_S and $M_S - M_{AF}$. The Resultant Achievement Motivation Test and the Prestatie Motivatie Test both are significantly ($p < .01$) related to total P_S ($r = -.38$ and $-.41$, respectively). Correlation of total P_S with the Prestatie Motivatie Test – Debilitating Anxiety subscale is slightly stronger ($r = -.48$). These findings imply a general tendency for persons who are high in M_S and particularly those who are also low in M_{AF} to overestimate P_S more than persons with the opposite predispositions.

In order to determine whether the biasing of P_S by M_S and M_{AF} differed across the P_S dimension, an objective-subjective P_S discrepancy score was analyzed via a 2×5 analysis of variance with repeated measures on the second factor (P_S). The discrepancy score was computed by summing the deviations between group-objective P_S and the subject's P_S rating for the seven problems in each of the five P_S categories and then adding 100 to each category sum. Thus, a score of 100 for a P_S category indicated that the subject's subjective P_S rating matched the group-objective difficulty of the items. Scores of less than 100 indicate overestimation of P_S, and larger scores indicate underestimation.

Analyses of variance of discrepancy scores were conducted using the Resultant Achievement Motivation Test and the Prestatie Motivatie Test – Debilitating Anxiety subscale. The results were nearly identical; those based on the latter measure of M_S are described below and are represented in Figure 2. The analysis of discrepancy scores yielded a highly significant main effect for item difficulty ($F = 238.91$, $df = 4/176$,

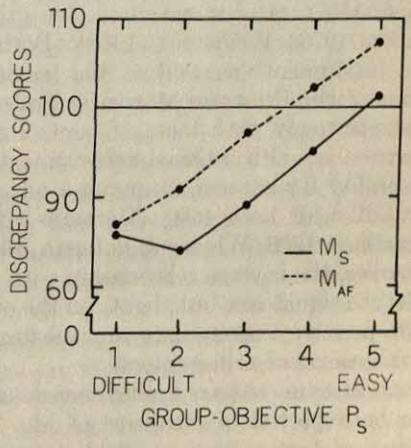


FIGURE 2. Discrepancy scores of motive to succeed (M_S) and motive to avoid failure (M_{AF}) subjects at different levels of probability of success (P_S). (Discrepancy scores over 100 indicate overestimation of P_S and scores under 100 indicate underestimation of P_S. The P_S levels are as follows: 1 = 0–10, 2 = 20–30, 3 = 45–55, 4 = 70–80, and 5 = 90–100.)

$p < .001$) in which overestimation of P_S was greatest when objective P_S was low. According to the same analysis, there was significantly less overestimation of objective P_S by M_{AF} than by M_S subjects ($F = 47.46$, $df = 1/44$, $p < .001$) and a significant Personality Type × P_S interaction ($F = 8.22$, $df = 4/176$, $p < .001$).

Both personality types tend to be most accurate when the task is easier, and most prone to overestimate when P_S is low, although the point of accuracy (100) is closer to the easier extreme of the P_S scale for M_S subjects. Moreover, M_{AF} subjects underestimate P_S on very easy tasks (P_S = 90–100) compared to M_S subjects ($t = 2.24$, $df = 44$, $p < .05$), while M_S persons overestimate P_S on moderately difficult ($P_S = .50$) and difficult tasks (P_S = 20–30) significantly more than M_{AF} persons ($t = 2.89$, $df = 44$, $p < .05$, and $t = 2.65$, $df = 44$, $p < .05$, respectively).

Discussion

In many respects, the Resultant Achievement Motivation Test (Mehrabian, 1969) and the Prestatie Motivatie Test (Hermans, 1970) perform similarly and satisfactorily with a junior high school population. The relationship between the two is relatively strong compared to correlations among objective M_S measures reviewed by others (Clarke, 1973; Weinstein, 1969). Both of these instruments as well as the resultant version of the Prestatie Motivatie Test correlate positively with locus of control and, in particular, with internal achievement responsibility for success, implying a measure of validity for both tests (Hermans, 1970; Mehrabian, 1968; Weiner & Potepan, 1970). Moreover, the inverse relationship with test anxiety is significant but slight; on the other hand, positive correlations are substantial with measures of achievement.

Both instruments are multidimensional as is to be expected of measures of M_S. The number of factors and variance accounted for by the first factor of the Resultant Achievement Motivation Test compare closely with Mehrabian's (1968) factor analysis of the earlier, 34-item version. The present factor analysis of the Prestatie Motivatie Test provided an opportunity to match Hermans' a priori factors with empirically derived factors. The success of the matching was clearly uneven. The Resultant Achievement Motivation Test is marred by relatively low internal consistency which, to our knowledge, had not been assessed previously. Mehrabian (1968) has reported higher reliability when stability over time was used as an estimate. The internal consistency of the Prestatie Motivatie Test obtained in the present study is high and slightly stronger than that reported by Hermans (1970). Moreover, the Prestatie Motivatie Test has the advantage of more stable subscales. Thus, the present findings related to the reliability and validity of the two objective measures are cause for guarded optimism regarding their use with school children, particularly in view of the inadequacies of other objective instruments (Weinstein, 1969).

The relationship between M_S and M_{AF} and P_S is not in complete accord with theoretical expectations (Figure 1), a point to be discussed later. Nevertheless, the closer approximation to expectations for the M_S persons when group standards of P_S are used is particularly interesting. Achievement motivation may be susceptible to social influences from gamelike tasks which generate social comparisons and competition (Maehr & Sjogen, 1971). In respect to the behavior of M_S subjects, at least, social influences may extend to the definitions of P_S and, accordingly, to I_S and I_F. Thus, P_S may be largely defined by standards of difficulty set by others, and I_S and I_F defined by pride or shame in accomplishments due to how the subject feels he is perceived in the eyes of others. An excellent ringtosser may feel considerable pride in making a shot which is relatively easy for him but markedly difficult for others, while a poor ringtosser may feel a great shame in missing a shot which is easy for others but difficult for him. The closer approximation of the model by M_S subjects using group definitions of P_S argues that social influences such as those described above are not extrinsic to achievement motivation but are an integral part of it, at least for ninth graders. The present findings with adolescents underscore Birney, Burdick, and Teevan's (1969) concern for a sharper delineation of individual and social

factors in determining P_S and suggest an investigation of developmental changes in the relative contribution each factor makes to the subjective definition of P_S.

Even when group standards are used to estimate P_S, the obtained curves (Figures 1a and 1b) depart from the theoretical model and from other findings (deCharms & Carpenter, 1968; Isaacson, 1964) in at least three respects: (*a*) M_{AF} subjects avoid difficult tasks; (*b*) M_{AF} subjects do not avoid moderately difficult tasks; and (*c*) M_S subjects' preference for $P_S = .50$ tasks is not as pronounced as expected. Unlike deCharms and Carpenter (1968), the procedures of this study did not suggest a gamelike task nor did they build in an inverse relationship between P_S and I_S. In addition, extrinsic factors such as the importance of grades may have been more at play in the present study. Even so, the failure of M_{AF} subjects to avoid moderately difficult tasks and to approach extremely difficult tasks has been observed on other occasions (Maehr & Sjogen, 1971). The clear-cut positive relationship between T_A and P_S for M_{AF} persons suggests that they may have treated the problems as matters of chance. The strong correlation between M_S and internal achievement responsibility for success, in particular, suggests this may be the case.

Atkinson and Feather (1966) view the influence of M_S and M_{AF} on P_S as a critical problem for achievement motivation theory. It is certainly an important issue in the application of that theory to educational problems. The significant, positive correlation between M_S and P_S confirms the lack of independence between these two factors when it comes to initial judgments of task difficulty (Atkinson, 1964, 1966; Brody, 1963; Moulton, 1965). The results of the analyses of variance also indicate greater overestimation of P_S by M_S than by M_{AF} subjects, an expected finding given the positive correlation between the two variables. But, contrary to Moulton's (1965) suggestion that M_{AF} subjects underestimate objective P_S, both groups overestimated.

The relationship between the two personality types and P_S is made more complex by the additional finding that biasing is not constant across levels of P_S. The biasing of P_S does not follow one trend suggested by Feather (1965), namely, that overestimation of P_S by M_S and underestimation of P_S by M_{AF} persons are greatest when achievement motivation is greatest, that is, at the $P_S = .50$ level. Overestimation of P_S by M_S persons is greatest for the more difficult problems which M_{AF} persons also overestimate. Brody (1963) proposed what amounts to a contrary relationship between M_S, M_{AF}, and P_S. He argued that M_S and M_{AF} persons distort their estimates of P_S to maximize their performance. If this is the case, M_S persons bias task difficulty so that all tasks are perceived as close to $P_S = .50$ as possible. Accordingly, they judge P_S most accurately when $P_S = 50$ and tend to overestimate P_S on difficult tasks and underestimate P_S on easy tasks. If we can assume that M_{AF} subjects also want to maximize resultant motivation as Brody suggests, they most accurately judge easy and difficult tasks and either underestimate or overestimate moderately difficult tasks. The present results are clearly inconsistent with this view. Overestimation of P_S by M_S subjects is relatively pronounced at $P_S = .50$ and least when P_S is high. For their part, M_{AF} persons are least accurate at the extreme P_S levels, underestimating P_S on easy problems and overestimating P_S on difficult ones. Their accuracy is greatest at an intermediate P_S level.

Finally, Feather (1965) proposed subjective attractiveness of success ($M_S \times I_S$) and subjective repulsiveness of failure ($M_{AF} \times I_F$) as constructs which may explain the influence of M_S and M_{AF} on P_S. The present results are most, but not entirely, consistent with this interpretation. The $M_S \times I_S$ formulation implies greater attractiveness of success and consequently greater overestimation of P_S with decreases in P_S and the accompanying increases in I_S. Thus, M_S persons are most accurate on the easiest tasks and most optimistic (i.e., overestimate P_S) on difficult tasks. This view seems to predict the P_S judgments of M_S persons, except on the extremely low P_S level where a slight deviation occurs (Figure 2).

Subjective repulsiveness of failure for M_{AF} subjects is greatest on easy tasks when both P_S and I_F are high, and it diminishes as tasks become more difficult. Accordingly,

M_{AF} persons are most accurate when the task is difficult and most pessimistic (i.e., underestimate P_S) when the task is easy. The present findings suggest such a relationship, but as the curves in Figure 2 indicate, the effect is not as strong as the hypothesis requires. Although M_{AF} subjects underestimate the P_S of easy tasks significantly more than M_S persons and M_S persons overestimate P_S in the difficult condition (P_S = 20–30) more than M_{AF} subjects, the most accurate estimate of P_S by M_{AF} subjects is not at the most difficult point on the P_S scale, as this view would predict. Nevertheless, in this important and relatively unexplored area of achievement motivation, the notions of subjective attractiveness of success and repulsiveness of failure seem to accommodate the present data best, or at least well enough to merit further investigation.

The findings of the present study suggest that application of achievement motivation theory to educational practice is hampered not only by the assessment of M_S, but perhaps even more so by problems associated with the estimation of P_S. The latter factor may possibly be subject to influences of M_S, of social factors, and of the developmental level of the learner, thereby making its manipulation by the teacher most difficult.

REFERENCES

Alpert, R., & Haber, R. N. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, 1960, **61**, 207–215.

Atkinson, J. W. *An introduction to motivation*. Princeton, N.J.: Van Nostrand, 1964.

Atkinson, J. W. The mainsprings of achievement-oriented activity. In J. D. Krumboltz (Ed.), *Learning and the educational process*. Chicago: Rand McNally, 1965.

Atkinson, J. W. Motivational determinants of risk-taking behavior. In J. W. Atkinson & N. T. Feather (Eds.), *A theory of achievement motivation*. New York: Wiley, 1966.

Atkinson, J. W., & Feather, N. T. Review and appraisal. In J. W. Atkinson & N. T. Feather (Eds.), *A theory of achievement motivation*. New York: Wiley, 1966.

Birney, R. C., Burdick, H., & Teevan, R. *Fear of failure*. New York: Van Nostrand-Reinhold, 1969.

Brody, N. N. Achievement, test anxiety, and subjective probability of success in risk-taking behavior. *Journal of Abnormal and Social Psychology*, 1963, **66**, 413–418.

Clarke, D. E. Measures of achievement and affiliation motivation. *Review of Educational Research*, 1973, **43**, 41–51.

Crandall, V. C., Katkovsky, W., & Crandall, V. J. Children's beliefs in their own control of reinforcements in intellectual-academic achievement situations. *Child Development*, 1965, **36**, 91–109.

deCharms, R., & Carpenter, V. Measuring motivation in culturally disadvantaged school children. In H. J. Klausmeier & G. T. O'Hearn (Eds.), *Research and development toward the improvement of education*. Madison, Wis.: Dembar Educational Research Services, 1968.

Farley, F. H., & Mealiea, W. L. Motivation and the recall of completed and incompleted achievement items. *Journal of Educational Research*, 1973, **66**, 302–306.

Feather, N. T. The relationship of expectancy of success to n achievement and text anxiety. *Journal of Personality and Social Psychology*, 1965, **1**, 118–126.

Hermans, H. J. M. A questionnaire measure of achievement motivation. *Journal of Applied Psychology*, 1970, **54**, 353–363.

Isaacson, R. L. Relation between n achievement, test anxiety, and curricular choices. *Journal of Abnormal and Social Psychology*, 1964, **68**, 447–452.

Maehr, M. L., & Sjogen, D. D. Atkinson's theory of achievement motivation: First step toward a theory of achievement motivation? *Review of Educational Research*, 1971, **41**, 143–161.

Mehrabian, A. Male and female scales of the tendency to achieve. *Educational and Psychological Measurement*, 1968, **28**, 495–502.

Mehrabian, A. Measures of achievement tendency. *Educational and Psychological Measurement*, 1969, **29**, 445–451.

Moulton, R. W. Effects of success and failure on level of aspiration as related to achievement motives. *Journal of Personality and Social Psychology*, 1965, **1**, 399–406.

Weiner, B. *Theories of motivation*. Chicago: Markham, 1972.

Weiner, B., Johnson, P. B., & Mehrabian, A. Achievement motivation and the recall of incompleted and completed exam questions. *Journal of Educational Psychology*, 1968, **59**, 181–185.

Weiner, B., & Potepan, P. Personality characteristics and affective reactions toward exams of superior and failing college students. *Journal of Educational Psychology*, 1970, **61**, 144–151.

Weinstein, M. S. Achievement motivation and risk preference. *Journal of Personality and Social Psychology*, 1969, **13**, 153–172.

Wolk, S., & DuCette, J. The moderating effect of locus of control in relation to achievement motivation variables. *Journal of Personality*, 1973, **41**, 59–70.

(Received October 24, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 609–613

COURSE EVALUATION: WHEN?[1]

NEIL A. CARRIER,[2] GEORGE S. HOWARD, AND WILLIAM G. MILLER
Southern Illinois University

Support was found for two hypotheses which declared that students attending the last regular meeting of a college course in introductory psychology give more favorable instructor and course evaluations than those attending the final examination only. Data pertinent to a third hypothesis suggested that, relative to the last-meeting evaluations, the final examination context has little or no effect on ratings.

Student evaluation of college courses and instructors has certainly increased in popularity in the last decade. Although Gustad (1967) reported a decline in its systematic use, possibly due to suspicion about its validity, pressures from students (Werdell, 1967) will probably see it increase during the seventies. The excellent review by Costin, Greenough, and Menges (1971) appears to support this prediction.

Probably the most common practice is to administer an evaluation form at the last regular meeting of a class. Administration at the final examination is another possibility but may be avoided for fear that the context will sour student attitudes.

However, a problem may exist with the first procedure when class attendance is not compulsory. It is implied by this question: How representative a sample of the total class population is the group of students who attend the last class meeting? These faithful attenders may have characteristics different than those of the less faithful nonattenders. If so, the omission of the latter (one hypothesis) may give spuriously favorable evaluations. With this in mind a simple study was designed.

The plan was to administer an evaluation form on the last day of class and again at the final examination. Accordingly, three data groups would result. First would be the group of students who filled out the evaluation form on the last day of class—the last-day-of-class group. Second would be the same students who completed the form at the final examination—the final-examination group. Third would be those students who completed the form at the final examination only—the final-examination-only group. By proper comparisons of data, three intuitively based hypotheses could be tested.

Hypothesis 1. A comparison of ratings from the last-day-of-class and final-examination-only groups shows that the last-day-of-class group provided the more favorable evaluation. Implicit here is the assumption that the last-day attenders are a more favorably disposed sample than are the final-examination-only attenders.

Hypothesis 2. A comparsion of ratings from the final-examination and final-examination-only groups shows that the final-examination group provided the more favorable evaluation. In effect, this comparison tests the same assumption as Hypothesis 1, using the same two samples of subjects, but has the possible advantage over the Hypothesis 1 check of controlling for the context in which the ratings were performed.

Hypothesis 3. A comparison of ratings from the last-day-of-class and final-examination groups shows no systematic depression of ratings in the final examination situation. Behind this hypothesis is the belief that the situational influence of the final ex-

[1] The authors wish to thank Mrs. Marilyn Perry for assistance in the preparation of data for analysis.

[2] Requests for reprints should be sent to Neil A. Carrier, Department of Psychology, Southern Illinois University, Carbondale, Illinois 62901.

TABLE 1

FIRST-ROOT VECTOR OF EACH OF THE 19 QUESTIONNAIRE ITEMS FOR GROUPS FINAL EXAMINATION AND FINAL EXAMINATION ONLY

Item	Correlation
1. Value of course	.48
2. Quality of lectures	.51
3. Lecturer's knowledge	.47
4. His interest	.68
5. Clarity of his explanations	.51
6. His voice quality	.13
7. His use of audiovisual aids	.42
8. Integration of course topics	.31
9. Coordination of lectures, readings, laboratory materials	.25
10. Quality of laboratory instructor	−.12
11. Value of laboratory sessions	−.11
12. Reading material interest	.25
13. Fairness of tests	.17
14. Course difficulty	.11
15. Intellectual stimulation from lectures	.47
16. Amount of reading	.26
17. Difficulty of reading	.33
18. Number of tests	−.01
19. Amount of studying done for course	.24

amination is not a very important contributor to ratings.

METHOD

Subjects and Final Examination

Subjects were students in a sophomore-level course in introductory psychology at the largest campus (approximately 20,000 students) of a midwestern state university. The final examination was a one-hour, 70-item multiple-choice test which was approximately 80% on material since the previous hourly and 20% on prior material.

Evaluation Form

The course and instructor evaluation form is one that was used for several years in the course. It consists of 19 items, the first 13 concerned with "qualitative aspects" and the remainder with "quantitative aspects." The form provides a frame of reference by asking the student to compare the course "with other similar lower division courses you have had." On the qualitative aspects, the choice is made from a 3-point scale bearing the labels "better," "about the same," and "poorer." The quantitative item labels are "more," "about the same," and "less." Table 1 gives shortened versions of the items.

Procedure

The form was administered at the last regular lecture meeting of two large lecture groups. To permit a comparison of these students' evaluations with those they would provide at the final examination, a coding system unique to each student was explained and used. Approximately 6% of the forms were unusable as a result of spoilage or of inconsistent code usage. The form was administered again one week later immediately prior to the final examination—obviously for the second time for some (Group Final Examination) and for the first time for others (Group Final Examination Only). Scoring was done by giving a scale value of 3 to a "better" or "more" response and a value of 1 for a "poorer" or "less" choice.

Because of the variety of scale topics, no simple summation of scores across items could be rationally done. Consequently, the hypothesized differences between the groups of data were investigated by means of a multiple discriminant analysis. The technique was employed as a way of characterizing group differences rather than as a means of classification, which is its more common use (Bock & Haggard, 1968).

RESULTS

The multiple discriminant analysis (Veldman, 1967) was performed comparing the ratings of Group Last Day of Class (n = 221) with the ratings of Group Final Examination Only (n = 174) to test Hypothesis 1. Analysis over the 19 items showed the groups to be significantly different (F = 2.97, df = 19/375, p = .0001). (The centroids of the discriminant space generated in this analysis were 2.894 for Group Last Day of Class and 2.499 for Group Final Examination Only.) Inspection of the data revealed that on 15 of the 19 items, Group Last Day of Class gave more favorable ratings than Group Final Examination Only, as predicted by Hypothesis 1.

A similar analysis comparing the final exam ratings of Group Final Examination (n = 221—the same students as Group Last Day of Class) and those of Group Final Examination Only (n = 174) also showed significant differences (F = 2.98, df = 19/375, p = .0001). (In this case, the centroids were 3.747 for Group Final Examination and 3.325 for Group Final Examination Only.) Inspection showed more favorable ratings by Group Final Examination than by Group Final Examination Only on 16 of the 19 items. Thus, Hypothesis 2 was also supported.

In view of the great similarity of these two comparisons, the remainder of the re-

sults deal only with the course evaluations completed on the day of the final examination—Groups Final Examination and Final Examination Only. The trace (sum of the roots) of the matrix formed as the ratio of the among-groups to the pooled, within-groups variance–covariance matrices was found to be .1512. The first root of this product matrix accounted for 99.98% of the trace. Since the trace is a measure of variance to be accounted for by the roots (where the roots satisfy the determinantal equation $|W^{-1}A - \lambda I| = 0$), one root sufficiently accounted for the available variance. This is, of course, the expected result when two groups are analyzed (Cooley & Lohnes, 1962).

Table 1 presents the loadings of each of the 19 questionnaire items with the first root's corresponding vector (discriminant function). As would be expected from the diverse coverage of the items, there are great differences in how heavily each item loads on the discriminant function.

Table 2 presents the results of the univariate F tests for the 19 items. In these comparisons, the between and within degrees of freedom are 1 and 393, respectively.

Finally, to shed light on the third hypothesis, Pearson correlations of test–retest reliability for the subjects in the last-day-of-class and final-examination groups were run for each questionnaire item. These are shown in Table 3. The range of these correlations is from .76 to .44. In view of the restricted range (1–3) of the item scales, these correlations are of reasonable magnitude. Also in Table 3 are the means on each item of Groups Last Day of Class and Final Examination. Inspection shows that although the Group Last Day of Class means are higher on 13 items and Group Final Examination means are higher on 5 (with 1 tie), the differences are generally miniscule and none approaches significance. Support for Hypothesis 3 seems indicated.

Discussion

In effect, this study was a replication of an unpublished one performed by one of the

TABLE 2

Results of the Univariate F Tests for the 19 Questionnaire Items for Groups Final Examination (FE) and Final Examination Only (FEO)

Item	Group FE		Group FEO		F
	M	*SD*	*M*	*SD*	
1. Value of course	2.64	.51	2.45	.58	12.47***
2. Quality of lectures	2.69	.49	2.48	.60	14.01***
3. Lecturer's knowledge	2.77	.43	2.61	.50	11.51***
4. His interest	2.79	.41	2.55	.55	25.21***
5. Clarity of his explanations	2.67	.47	2.47	.60	13.66***
6. His voice quality	2.58	.55	2.53	.59	.91
7. His use of audiovisual aids	2.78	.43	2.64	.52	9.17**
8. Integration of course topics	2.51	.52	2.39	.58	4.87*
9. Coordination of lectures, readings, laboratory materials	2.47	.57	2.36	.61	3.33
10. Quality of laboratory instructor	2.43	.67	2.48	.59	.79
11. Value of laboratory sessions	2.30	.65	2.35	.66	.61
12. Reading material interest	2.31	.62	2.20	.66	3.29
13. Fairness of tests	2.23	.61	2.15	.60	1.56
14. Course difficulty	2.61	.52	2.57	.53	.62
15. Intellectual stimulation from lectures	2.52	.56	2.32	.61	11.50***
16. Amount of reading	2.70	.48	2.60	.57	3.48
17. Difficulty of reading	2.45	.51	2.32	.58	5.83*
18. Number of tests	2.33	.56	2.33	.52	.01
19. Amount of studying done for course	2.63	.55	2.53	.61	2.89

* $p < .05$.
** $p < .01$.
*** $p < .001$.

TABLE 3

MEANS AND PEARSON CORRELATION RELIABILITY COEFFICIENTS FOR GROUPS LAST DAY OF CLASS (LD) AND FINAL EXAMINATION (FE) ON THE 19 QUESTIONNAIRE ITEMS

Item	Group LD	Group FE	F^a	r
1. Value of course	2.61	2.64	.55	.70
2. Quality of lectures	2.71	2.69	.15	.67
3. Lecturer's knowledge	2.78	2.77	.05	.57
4. His interest	2.76	2.79	.80	.53
5. Clarity of his explanations	2.64	2.67	.45	.44
6. His voice quality	2.61	2.58	.27	.60
7. His use of audiovisual aids	2.83	2.78	1.36	.52
8. Integration of course topics	2.43	2.51	1.97	.52
9. Coordination of lectures, readings, laboratory materials	2.46	2.47	.06	.45
10. Quality of laboratory instructor	2.43	2.43	.01	.73
11. Value of laboratory sessions	2.26	2.30	.40	.66
12. Reading material interest	2.22	2.31	2.34	.48
13. Fairness of tests	2.19	2.23	.29	.55
14. Course difficulty	2.56	2.61	.94	.66
15. Intellectual stimulation from lectures	2.53	2.52	.07	.59
16. Amount of reading	2.65	2.70	1.40	.56
17. Difficulty of reading	2.38	2.45	1.62	.56
18. Number of tests	2.27	2.33	1.41	.50
19. Amount of studying done for course	2.60	2.63	.26	.76

[a] None of the F values in this column approaches significance.

authors a year previously. Although more appropriate statistical procedures were employed in the present effort, similar findings are apparent. The confirmations of Hypotheses 1 and 2 suggest that in a course with voluntary attendance the last-day attenders are more disposed to give favorable evaluations than the group that attends the final examination only. The implications of this, for determining which procedure to follow in obtaining the most favorable evaluations, now present a challenge to the instructor's conscience.

The evidence pertinent to Hypothesis 3 supports the contention that the final-examination context does not have an appreciable depression effect on ratings. The implication of this for the instructor who *also* wishes to get the most representative sample of student reactions is to give evaluation forms at the final examination with relative impunity.

However, a few caveats are in order. The comparison of Group Final Examination and Group Final Examination Only data (Hypothesis 2) was performed to control for the possibility that the differences between Group Last Day of Class and Group Final Examination Only data (Hypothesis 1) could have been due to negative feeling induced by the testing situation. This control may not have been completely effective. It might be argued that student responses in the final-examination situation were made with an effort on their part to be consistent with the last-day-of-class responses which they made a week before. A replication of this study to remove this possible influence could easily be done: simply record by some coding system the students attending the last class period but don't administer the evaluation form, then give the form at the final examination to all present and obtain the coded information again. Analysis would then contrast the evaluations of students who had attended the last class meeting with those of students who had not. Any possible consistency bias would thereby be avoided. The test–retest reliability information would be lost, however.

A second variable of possible significance to Hypothesis 3 might be the degree of difficulty of the final examination. As noted, the examination here was rather minimally comprehensive and only one hour in length. Would the prospect of facing a more difficult

comprehensive examination of two or three hours duration depress ratings substantially more than was apparent here?

Finally, a further refinement of research on evaluations would be to partial out variables that previous research suggests may have potency (e.g., expected grades: Costin, Greenough, & Menges, 1971) before contrasting the centroids of the treatment groups.

REFERENCES

Bock, D. R., & Haggard, E. The use of multivariate analysis of variance in behavioral research. In D. Whitla (Ed.), *Handbook of measurement and assessment in behavioral science*. Reading, Mass.: Addison-Wesley, 1968.

Cooley, W. W., & Lohnes, P. R. *Multivariate procedures for the behavioral sciences*. New York: Wiley, 1962.

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, 1971, **41**, 511–535.

Gustad, J. W. Evaluation of teaching performance: Issues and possibilities. In C. B. T. Lee (Ed.), *Improving college teaching*. Washington, D.C.: American Council on Education, 1967.

Veldman, D. J. *Fortran programming for the behavioral sciences*. New York: Holt, Rinehart & Winston, 1967.

Werdell, P. R. *Course and teacher evaluation*. (2nd ed.) Washington, D.C.: United States National Student Association, 1967.

(Received October 29, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 4, 614-622

EFFECTS OF INSTRUCTIONAL OBJECTIVES USED AS ORIENTING STIMULI OR AS SUMMARY/REVIEW UPON PROSE LEARNING[1]

ROBERT KAPLAN[2]
Bell Laboratories,
New Brunswick, New Jersey

FRANCINE G. SIMMONS
Temple University

Instructional objectives were located before a text as orienting stimuli for selective attention or after a text as a summary/review. Four experimental treatments were examined: objectives presented before text (*a*) with and (*b*) without relevant information and objectives presented after text (*c*) with and (*d*) without relevant information. Three processes (repetition, selection, and search) were hypothesized to account for the results of the experimental treatments. The main finding was that performance on information relevant to an objective was relatively high whether the objectives were presented before or after the text. Performance on incidental material was greater for objectives located after the text than before the text. More inspection time was consumed by the experimental groups than by a group that received no objectives.

The use of instructional objectives to direct learners to relevant portions of a text has been shown to enhance objective-relevant (intentional) learning (Duchastel & Merrill, 1973, for a partial review; Kaplan, 1973; Kaplan & Rothkopf, 1974; Rothkopf & Kaplan, 1972). The objectives typically used in these studies direct the learner to relevant material but do not repeat the relevant information. This is similar to providing the learner with unanswered questions. In both cases, the learner then must search the text for the information relevant to the objective or question. Frase (1968, 1970) suggested that questions located prior to text segments serve as orienting stimuli which influence the learners' inspection of the text. Frase, and later Patrick (see Frase, 1973) and Boyd (1973), suggested that questions used in this way might be the occasion for selective attention to relevant text items, namely, incidental learning might be depressed with prequestions under some conditions.

Other studies investigating the use of objectives or questions after text segments suggested that they may function as a summary or review rather than as orienting stimuli (Bruning, 1968; Frase, 1967; Rothkopf, 1966). These studies also showed that intentional learning was greater when the answers to questions were provided than when no answers were given. In general, overall retention tended to be greater when questions were located after rather than before text segments. The distinguishing feature of these studies is that the objectives or questions presented after text segments could not be used as orienting stimuli which permit the learner to selectively attend to specific information during inspection of the text. Frase (1967) suggested that increased performance under these conditions could result from several factors: (*a*) review, (*b*) repetition of relevant material, and (*c*) practicing test-like events.

The relative effects of objectives or questions located before or after a text are less clear with respect to incidental (nonobjective-relevant) learning. Several studies have

[1] We are greatly indebted to E. M. Burgin for her valuable assistance throughout the experiment.
[2] Requests for reprints should be sent to Robert Kaplan, Bell Laboratories, CPI L-101, P.O. Box 2020, New Brunswick, New Jersey 08903.

TABLE 1

HYPOTHESIZED STRATEGY, CONSEQUENCE, AND OUTCOME OF TREATMENTS

Objective treatment	Hypothetical strategy	Suggested consequence			Predicted outcome[a]	
		Repetition	Selection	Search	Intentional	Incidental
No objective (reference group)	read text nonselectively	no	no	no	=[b]	
Before/without information	read objectives; search text selectively for answers	partial	yes	yes	>	<
Before/with information	read objectives with information; ignore text	no	yes	no	=	<
After/with information	read text nonselectively; read objectives with information	yes	no	no	>	=
After/without information	read text nonselectively; read objectives without information	partial	no	no	=	=

[a] Comparison between treatment and reference groups.
[b] Comparison between test item sets for reference group.

reported greater incidental learning for groups receiving objectives or questions than for nonobjective control groups (Bruning, 1968; Frase, 1967; Rothkopf, 1966; Rothkopf & Bisbicos, 1967). Frase (1967) and Rothkopf (1966) both found greater incidental learning when questions were presented after rather than before text segments. The previously mentioned Kaplan and Rothkopf (1974) and Rothkopf and Kaplan (1972) studies found varied effects for incidental learning when objectives were presented prior to text.

Based upon the results of all the studies previously discussed, it would seem that at least three processes influence the subjects' learning with objectives (repetition, selection, and search). The primary purpose of the present study was to use these constructs to predict learning outcomes. This was done by determining the learning effects of objectives presented prior to or after a text when the objectives were written with or without relevant information. An objective with information is analogous to a question containing an answer, while an objective without information is analogous to a question not containing an answer. It is possible, then, to hypothesize specific inspection strategies that subjects may adopt for each of the four treatments and for a reference group receiving no objectives (Table 1).[3] The three hypothetical processes that allow these predictions are repetition, selection, and search. Each strategy is associated with a suggested consequence in terms of the processes employed and the subsequent predicted outcomes. It is assumed that (*a*) repetition will result in increased learning of the repeated information and in increased inspection time; (*b*) selection will result in increased learning of selected information (because there is less information to attend to), decreased learning of information not selected (incidental learning), and in a reduction of inspection time; and (*c*) search will result in increased learning of information not selected and in increased inspection time, both relative to selection alone.

The reference group is hypothesized to use a strategy that gives equal attention to each text sentence (i.e., nonselective text inspection). The suggested consequence of this strategy is that the subject would not have the opportunity for repetition (of objective and then text sentence), selection (of only relevant material for study), or search (for relevant sentences within the text). The predicted outcome would be equal perform-

[3] Appreciation is extended to L. T. Frase for his contribution to the information in Tables 1 and 2.

ance on material that was used to measure intentional and incidental learning for the treatment groups (Table 1). When objectives without relevant information are presented prior to text (i.e., available concurrently with the text) and the subject is instructed that the test will only pertain to objective-relevant material, an efficient strategy would be for the subject to read the objectives and then search the text to select relevant information. The consequence of this strategy would involve partial repetition (the objective is repeated as part of the text sentence), search (searching the text for relevant sentences), and selection (focusing attention on objective-relevant text sentences). Based upon these consequences, the treatment group's intentional learning is predicted to be greater than the reference group's performance on the same test items. The treatment groups' lack of attention to nonobjective-relevant material and the reference group's equal attention to this material is the basis for predicting treatment group incidental performance to be less than that of the reference group (Table 1). When objectives with relevant information are presented prior to text, the strategy of attending to the objectives but not to the text may be most efficient. The consequence of this strategy would involve the process of selection (i.e., select objectives only for study) but not the process of repetition (i.e., no repetition of objective and then text sentence) or search (i.e., no text search). Although the treatment group's use of selection reduces the amount of material attended to, the difference was not considered large enough to produce significant learning differences. Further, neither the treatment nor the reference group benefits from repetition or search. Therefore, the treatment group's performance on objective-relevant material is predicted to be about equal to that of the reference group. Conversely, the treatment group's nonobjective-relevant learning is predicted to be less than the reference group's because of the treatment group's lack of attention to this material (Table 1). When objectives with information are presented after text inspection (not concurrently), the text must be read nonselectively (same as control group). Then the objectives with information are read as a summary/review of relevant information. The suggested consequence of this strategy is the use of repetition but not selection or search. Due to the repetition advantage, the treatment group is predicted to learn more objective-relevant material than the reference group. Because both groups read the nonobjective-relevant material nonselectively, they are predicted to perform about equally on this material (Table 1). Finally, when objectives without information are presented after text inspection, the text is again read nonselectively, and the subsequent objectives without information provide only a partial repetition of relevant text sentences, one which is not sufficient to produce learning. Neither selection nor search is available for this treatment. Hence, the predicted outcome is for equal performance on intentional and incidental items for the treatment and reference groups (Table 1).

Several predictions can be made for the amount of inspection time used for each of the above treatments. Objectives presented before text are predicted to result in less inspection time than objectives presented after text. This prediction is made because the objectives-after-text treatments require time, first, to read the text nonselectively and, then, require additional time for the second task of inspecting the objectives as a summary/review. Objectives before text are predicted to result in slightly greater inspection time when they are presented without than with information. This prediction is based upon the additional search time required when the objectives do not contain information. Conversely, objectives after text are predicted to result in slightly greater inspection time when they are presented with than without information. This prediction was made because objectives serve as a better summary/review when they are presented with than without information; thus, the learner is expected to spend more time using the objectives with information. Finally, all of the treatment groups are predicted to result in greater inspection time than the reference group, with

the exception of the objectives-with-information-before-text group. This is predicted because all of the former treatments have the combined task of inspecting objectives and text. However, the latter treatment group inspects objectives only, thus they have to attend to less information than the reference group.

A more specific goal of the present experiment, then, is to test these predictions.

Method

Materials

The two experimental passages were used in previous studies (Kaplan, 1973; Kaplan & Rothkopf, 1974; Rothkopf & Kaplan, 1972). The passages were segments of two textbooks prepared by the Systems Training Department, Bell Laboratories, Piscataway, New Jersey.[4] The content of the passages pertained to printer specifications for designing forms (Passage 1) and to an introduction to business information systems (Passage 2) and comprised 842 and 1,091 ($\bar{X}$ = 967) words in 60 and 54 ($\bar{X}$ = 57) sentences, respectively.

Two types of objectives were prepared for this study (without and with information). Objectives without information were the same as the specifically phrased objectives used in the above mentioned Kaplan and Rothkopf studies. A specifically phrased objective was defined as having been empirically determined to require the subject to learn only one passage sentence (for details, see Rothkopf & Kaplan, 1972). The distinguishing feature of this type of objective was that it did not contain relevant text information; that is, the objective told the subject what to learn but did not reveal the information necessary to answer a test question about that objective. An example of an objective without information is, "Learn how many years 75% to 100% rag paper will last." Conversely, a matching set of objectives was prepared that did contain the relevant text information necessary to answer test questions about those objectives. An example of an objective with information that corresponds to the previous example is, "Learn that 75% to 100% rag paper will last 30 to 40 years." The number of matching objectives, with and without information, for both passages, was 36 and 33 ($\bar{X}$ = 34.5). Thus, the proportion of objective-relevant passage sentences (34.5) to total passage sentences (57) was about 60%; that is, the subjects were instructed to learn 60% of the passage, which was equivalent to "density-60%" in the previous studies.

Both types of objectives were located either before or after the passage. The only difference was in the initial instruction which stated that the subject should learn (before) or should have learned (after) the list of objectives.

The tests were also the same as those used in the previous studies. They consisted of short-answer, fill-in-the-blank items. A test question was written for almost every sentence in Passages 1 and 2 (56 and 52 questions, respectively). This procedure permitted measuring both objective-relevant (intentional) learning and nonobjective-relevant (incidental) learning. The test questions for each passage were presented in three different random orders.

Procedure

All subjects received a set of sequentially numbered manila envelopes containing experimental materials. Each envelope had preliminary instructions printed on the outside and specific instructions printed on pink paper inside the envelope. All subjects proceeded at their own pace without time limitations. Those subjects receiving objectives before the passage, either with or without information, were given three envelopes consisting of (*a*) objectives and a passage, (*b*) a test, and (*c*) supplementary reading material to occupy subjects finishing early. These subjects were told that they would be tested only on information in the passage relevant to the objectives. By testing them on additional nonobjective-relevant items, a measure of incidental learning was achieved. Those subjects receiving objectives after the passage, either with or without information, were given four envelopes consisting of (*a*) a passage, (*b*) objectives, (*c*) a test, and (*d*) supplementary material. These subjects were given the general instruction in Envelope 1 to learn everything in the passage. They were then given the objectives in Envelope 2 as a review of important material. The reference group receiving no objectives was given three envelopes containing (*a*) a passage, (*b*) a test, and (*c*) supplementary material. However, they were only given the traditional instruction to learn everything in the passage. The subjects participated in groups of no less than 30 in a session. They were required to record their start and stop inspection times for each envelope. A digital clock with 2 × 5 inch numerals was provided for that purpose.

Design

Percentage correct responses were analyzed with a 2 × 2 × 2 analysis of variance with (*a*) two levels of objective information (with and without), (*b*) two levels of objective location (before and after), and (*c*) two levels of learning (intentional and incidental) with repeated measures on the last factor. In addition, a 2 × 2 analysis of variance was performed on log transformations of inspection time with (*a*) two levels of information and (*b*) two levels of location. Inspection time data consisted

[4] Gratitude is extended to F. L. Stevenson, Head, Systems Training Department at Bell Laboratories for permitting us to use the experimental materials.

of the combined time spent with the passage and objectives.

Subjects

The subjects were tenth-, eleventh-, and twelfth-grade students from three New Jersey high schools. Half of the subjects were paid volunteers who participated after the last school class. The other half of the subjects participated during school hours and were aware that a contribution to the student activity fund was being made. Subjects from both schools were randomly assigned to all treatments. The subjects consisted of 118 males and 182 females ($N = 300$). This allowed for 60 subjects in each of the four treatment groups and the reference group.

Results

Intentional and Incidental Learning

Two main effects of this analysis were not significant. Specifically, no differences in performance were found (*a*) between objectives located prior to ($\bar{X} = .42$) and after ($\bar{X} = .46$) the passage ($F = 3.26$, $df = 1/236$, $p > .05$) or (*b*) between objectives containing information ($\bar{X} = .44$) and not containing information ($\bar{X} = .44$; $F = .06$, $df = 1/236$, $p > .05$). However, Figure 1 shows a significant Location of Objectives × Information Content interaction ($F = 10.04$, $df = 1/236, p < .01$).

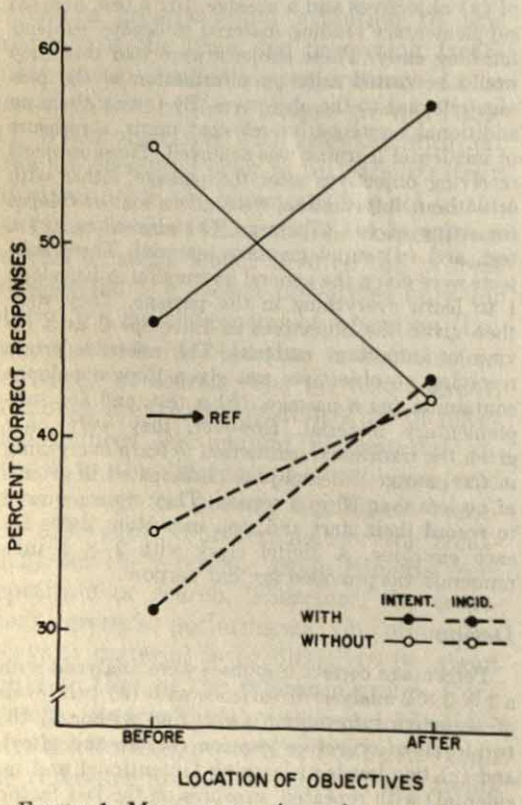


Figure 1. Mean proportion of correct intentional (intent.) and incidental (incid.) test items for treatment groups receiving objectives with and without relevant information as a function of objective location and for a reference (REF) group.

As expected, performance on intentional items ($\bar{X} = .50$) was greater than that for incidental items ($\bar{X} = .38$; $F = 81.03$, $df = 1/236$, $p < .001$). A significant interaction was found between the location of objectives and learning ($F = 16.06$, $df = 1/236$, $p < .001$). This interaction was a result of increases in incidental learning when objectives were presented after ($\bar{X} = .43$) rather than before ($\bar{X} = .33$) the passage, while no differences occurred for intentional learning. The Objective Information Content × Learning interaction was not significant ($F = 2.62$, $df = 1/236$, $p > .05$). Finally, the three-way Location × Information Content × Learning interaction was significant ($F = 13.60$, $df = 1/236$, $p < .001$). This interaction can be seen in Figure 1 as resulting from the Location × Information interaction found for intentional learning and the simple increase in incidental learning with the after-passage objective location.

Paired comparisons between data points in Figure 1 were accomplished with the Newman-Keuls technique. For intentional learning, objectives without information resulted in greater performance when located prior to the passage ($\bar{X} = .56$) than after the passage ($\bar{X} = .42$; $q = 5.04$, $df = 236$, $r = 2, p < .01$). Conversely, objectives with information resulted in greater performance when located after the passage ($\bar{X} = .57$) than prior to the passage ($\bar{X} = .46$; $q = 4.15$, $df = 236, r = 2, p < .01$). In addition, performance on objectives without information was greater than objectives with information when located prior to the passage, while the reverse was found when objectives were located after the passage ($q = 3.59$, $df = 236$, $r = 2$, $p < .05$ and $q = 5.59$, $df = 236, r = 2, p < .01$, respectively). For incidental learning, performance on objectives with information was greater when

located after the passage ($\bar{X} = .43$) than before the passage ($\bar{X} = .31$; $q = 4.44$, $df = 236$, $r = 2$, $p < .01$). No difference in performance was found between objectives without information when located after the passage ($\bar{X} = .42$) and when located before the passage ($\bar{X} = .35$; $q = 2.70$, $df = 236$, $r = 2, p > .05$).

Comparisons between the reference group receiving no objectives and the treatment groups were made with one-tailed Dunnett tests. The error terms were derived from two separate analyses of variance for intentional and incidental learning items. Both analyses were single factors with five levels of groups (before/with, before/without, after/with, after/without, and reference). The intentional learning comparisons showed that objectives located before the passage without information and after the passage with information resulted in greater performance than the reference group ($t = 3.85$ and 4.21, respectively, $df = 295$, $k = 5$, $p < .01$). No difference was found between the reference group and objectives located before the passage with information or after the passage without information ($t = 1.47$ and .49, respectively, $df = 295$, $k = 5$, $p > .05$) for both tests. The incidental learning comparisons showed that the reference group performance was greater than objectives located before the passage both with and without information ($t = 3.76$, $p < .01$ and $t = 2.48$, $p < .05$, respectively, $df = 295$, $k = 5$). No differences were found between the reference group and objectives located after the passage either with or without information ($t = .10$ and .13, respectively, $df = 295$, $k = 5$, $p > .05$).

Inspection Time

Figure 2 shows that less inspection time was consumed when objectives were located before the passage ($\bar{X} = 16.66$) than when they were located after the passage ($\bar{X} = 18.47$; $F = 5.29$, $df = 1/216$, $p < .05$). Approximately the same amount of time was consumed by groups receiving objectives with ($\bar{X} = 17.42$) and without ($\bar{X} = 17.71$) information ($F = .22$, $df = 1/216$, $p > .05$). The interaction between these factors was not significant ($F = 1.75$, $df = 1/216$, $p > .05$). Although this interaction was not significant, paired comparisons using the Newman-Keuls technique showed that for objectives with information, more time was consumed when located after ($\bar{X} = 18.90$) than when located prior ($\bar{X} = 15.93$) to the passage ($q = 3.64$, $df = 216$, $r = 2$, $p < .01$). Conversely, objectives without information required about the same inspection time when located before ($\bar{X} = 17.39$) or after ($\bar{X} = 18.03$) the passage ($q = .96$, $df = 216, r = 2, p > .05$).

Inspection time comparisons between the reference group and the treatment groups were made with one-tailed Dunnett tests. The error term for this analysis was derived from a single-factor analysis of variance on the log inspection time data, with five levels of groups (four treatment groups and the reference group). The reference group ($\bar{X} = 11.71$) consumed less inspection time than (*a*) objectives before/with information, (*b*) objectives before/without information, (*c*) objectives after/with information, and (*d*) objectives after/without information ($t = 3.48$, 4.80, 6.15, and 5.50, respectively, $df = 274$, $k = 5$, $p < .01$) for all four tests.

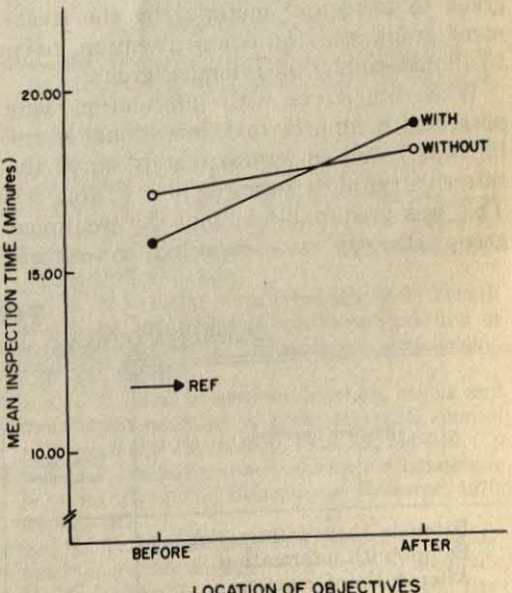


FIGURE 2. Mean inspection time for treatment groups receiving objectives with and without relevant information as a function of objective location and for a reference (REF) group.

Discussion

The results of this experiment tend to confirm the predictions posited in the introduction except for the inspection times between objectives with information located before text and the reference group. Table 2 shows the predictions and subsequent scores obtained for the treatment and control groups for intentional and incidental learning. The increased intentional learning found when objectives without information were presented prior to text over the reference group's performance on the same items was predictable from previous studies (Kaplan, 1973; Kaplan & Rothkopf, 1974; Rothkopf & Kaplan, 1972). This finding may be understood in terms of Frase's (1968, 1970) suggestion that questions or objectives of this type serve as orienting stimuli that result in selective attention to relevant material. When objectives are used in this way, an opportunity exists for partial repetition, selection, and search (Table 1). Conversely, this treatment group's incidental learning was predicted and found to be less than the reference group's performance on the same items (Table 2). This difference is attributed to the lack of attention given to incidental material by the treatment group and the equal attention given to all material by the reference group.

When objectives with information were presented before the text, intentional learning was found to approximately equal the reference group's performance (Table 2). This was predictable in that the treatment group strategy was expected to exclude repetition and search (Table 1). The slightly greater treatment group mean performance ($\bar{X} = .45$) over the reference group ($\bar{X} = .40$) was probably due to the reduction of material attended to by the treatment group as a result of selection (i.e., attending to the objectives only). As expected with this strategy, the treatment group's incidental learning was less than the reference group performance because the treatment group would not have attended to incidental text material.

Intentional learning was greater for subjects receiving objectives with information after text than for the reference group. This was predictable in that the treatment group could benefit from repetition and summary/review (Table 2). Incidental learning was found to be about equal for both groups. This was predicted because both groups read the text nonselectively.

Performance, when objectives without information were presented after the text, was found to be the same as the reference group for intentional and incidental learning (Table 2). This was predicted because the text was read nonselectively by both groups and the objectives served as an inadequate summary/review in that relevant information was excluded.

These results suggest that the increased learning found with objectives resulted from more than one process (repetition, selection, and search). Further, the processes employed seem to vary with specific learning strategies which are influenced by the type and method of presenting objectives. Thus,

TABLE 2

Predicted Outcome and Results of Treatments

Objective treatment	Intentional			Incidental		
	Predicted outcome[a]	$\bar{X}$ Score		Predicted outcome[a]	$\bar{X}$ Score	
		Reference	Treatment		Reference	Treatment
Before/without information	>	.40	.55	<	.43	.35
Before/with information	=	.40	.45	<	.43	.31
After/with information	>	.40	.58	=	.43	.42
After/without information	=	.40	.42	=	.43	.41

[a] Comparison between treatment and reference groups.

the results of several studies previously mentioned can be explained in terms of these processes.

The inspection time analysis shows that less time is consumed by treatments receiving objectives prior to text than after text (Figure 2). This may be due to the procedures employed in the present experiment. All subjects were permitted to inspect the experimental materials at their own rate. Those subjects receiving objectives before the text actually inspected the objectives and text concurrently. Conversely, subjects receiving objectives after the text inspected the text first and upon completion of that task they were given the objectives. These subjects were not instructed that the objectives would follow the text. Therefore, they may have decided to inspect the text for a longer time than the objectives-before group in that they would assume that this was their only opportunity to learn the material. Subsequent inspection of the objectives then required additional time. In addition, the objectives before text treatments could employ a selective attention strategy to reduce the amount of material studied, thus, reducing their inspection times. As predicted, objectives without information located before the text and both treatments with objectives located after the text resulted in greater inspection time than the reference group. This was due to the additional time required by these treatment groups to use the objectives. However, the one unpredicted finding was the increased time taken with objectives located before the text with information over the reference group. This treatment was expected to use less time than the reference group because they could eliminate all incidental material. Perhaps the prediction would have been substantiated if fewer objectives were presented. That is, even if only the objectives were attended to, they represented 60% of the total number of text sentences. Thus, there was a difference of only 23 sentences between treatment and reference groups. In addition, the treatment group probably used some time to initially inspect the objectives and text before deciding to adopt the strategy of attending to objectives only. Perhaps more important, all treatment groups consumed more inspection time than the reference group (Figure 2). For practical purposes, this finding must be considered with respect to performance. That is, while more time is required for the before/without and the after/with relevant information treatments than the reference group, these two treatment groups' performance is greater than that of the reference group. If mastery of material is important, then the use of objectives may be desirable even though more inspection time is consumed.

The results of this experiment suggest an obvious future investigation. The combination of objectives without relevant information before the text and objectives with relevant information after the text may increase objective-relevant performance over either treatment alone. This would result because of the combined advantage of search and selection when objectives are presented before the text and repetition when located after the text. Particular attention should be given to the effects upon incidental learning and inspection time under this treatment.

REFERENCES

Boyd, W. M. Repeating questions in prose learning. *Journal of Educational Psychology*, 1973, **64**, 31–38.

Bruning, R. H. Effects of review and testlike events within the learning of prose materials. *Journal of Educational Psychology*, 1968, **59**, 16–19.

Duchastel, P. C., & Merrill, P. F. The effects of behavioral objectives on learning: A review of empirical studies. *Review of Educational Research*, 1973, **43**, 53–69.

Frase, L. T. Learning from prose material: Length of passage, knowledge of results and position of question. *Journal of Educational Psychology*, 1967, **58**, 266–272.

Frase, L. T. Effect of question location, pacing and mode upon retention of prose material. *Journal of Educational Psychology*, 1968, **59**, 244–249.

Frase, L. T. Boundary conditions for mathemagenic behavior. *Review of Educational Research*, 1970, **40**, 337–347.

Frase, L. T. Maintenance and control in the acquisition of knowledge from written materials. In R. O. Freedle & J. B. Carroll (Eds.), *Language comprehension and the acquisition of knowledge*. Washington, D.C.: V. H. Winston, 1973.

Kaplan, R. The effects of learning prose with part vs whole presentations of instructional objectives. Paper presented at the Annual Convention of the American Psychological Association, Montreal, September 1973.

Kaplan, R., & Rothkopf, E. Z. Instructional objectives as directions to learners: Effects of passage length and amount of objective-relevant content. *Journal of Educational Psychology*, 1974, **66**, 448–456.

Rothkopf, E. Z. Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, 1966, **3**, 241–249.

Rothkopf, E. Z., & Bisbicos, E. E. Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology*, 1967, **58**, 36–61.

Rothkopf, E. Z., & Kaplan, R. Exploration of the effect of density and specificity of instructional objectives on learning from text. *Journal of Educational Psychology*, 1972, **63**, 295–302.

(Received November 2, 1973)

Manuscripts Accepted for Publication in the *Journal of Educational Psychology*

Acquisition Process and Resilience under Varying Testing Conditions for Structurally Different Problem-Solving Procedures. Richard E. Mayer (Department of Psychology, Indiana University, Bloomington, Indiana 47401).

Acquisition and Nonspecific Transfer Effects in Prose Learning as a Function of Question Form. James F. Voss (Department of Psychology, University of Pittsburgh Pittsburgh, Pennsylvania 15260).

Longitudinal Changes in Intellectual Functioning in Preschool Children: Schooling and Age-Related Effects. L. R. Goulet (College of Education, University of Illinois, Urbana, Illinois 61801), Kerry G. Williams, and Carl M. Hay.

Effects of Cognitive Intervention on Latent Ability Measured from Analogy Items. (Susan E. Whitely, University Measurement Services Center, University of Minnesota, 9 Clarence Avenue S.E., Minneapolis, Minnesota 55414).

Student Teams and Learning Games: Their Effects on Cross-Race and Cross-Sex Interaction. David L. Devries (Center for Social Organization of Schools, The Johns Hopkins University, Baltimore, Maryland 21218) and Keith J. Edwards.

Imagery Instructions and Drawings in Learning Prose. Ronald W. Rasco (Computer Applications Laboratory, 1-A Tully Building, Florida State University, Tallahassee, Florida 32306), Robert D. Tennyson, and Richard C. Boutwell.

Effect of Prior Teaching Evaluations and Lecture Presentation on Ratings of Teaching Performance. Raymond P. Perry (Department of Psychology, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada), R. Ronald Niemi, and Keith Jones.

Effect of Intent-Action Discrepancy and Student Performance Feedback on Teacher Behavior Change. Walter Doyle (College of Education, North Texas State University, Denton, Texas 76203) and Judith McNally Redwine.

Cognitive Results Based on Different Ages of Entry to School: A Comparative Study. Gilbert R. Austin (Bureau of Educational Research and Field Service, University of Maryland, College Park, Maryland 20742) and T. Neville Postlethwaite.

Effects of a Four-Year Follow Through Program on Economically Disadvantaged Children. Willa D. Abelson (Child Study Center, Yale University, New Haven, Connecticut 06510), Edward Zigler, and Cheryl L. DeBlasi.

Abstractive Processes in the Remembering of Prose. Ronald E. Johnson (SCC "G," Purdue University, West Lafayette, Indiana 47907).

Indirect Review and Priming through Questions. Ernst Z. Rothkopf (Learning and Instruction Research Department, Bell Laboratories, 600 Mountain Avenue, Murry Hill, New Jersey 07974) and Marjorie J. Billington.

Encoding Speed, Rehearsal, and Probed Recall of Dyslexic Boys. Carl Spring (Department of Education, College of Letters and Science, University of California, Davis, California 95616) and Carolyn Capps.

Effects of Learning Prose with Part versus Whole Presentations of Instructional Objectives. Robert Kaplan (Bell Telephone Laboratories, CPI L101, P.O. Box 2020, New Brunswick, New Jersey 08903).

Sex Differences in Comprehension of High- and Low-Interest Reading Material. Steven R. Asher (Department of Educational Psychology, University of Illinois, Urbana, Illinois 61801) and Richard A. Markell.

Journal of Educational Psychology
1974, Vol. 66, No. 4, 623-630

TRAINING FOR THEMATIC-FANTASY PLAY IN CULTURALLY DISADVANTAGED CHILDREN: PRELIMINARY RESULTS[1]

ELI SALTZ[2] AND JAMES JOHNSON

Center for the Study of Cognitive Processes, Wayne State University

Reported herein are the preliminary findings of a broader longitudinal study investigating the effects of fantasy play intervention on socially and economically disadvantaged preschoolers. Young children directed in the role enactment of imaginary stories were found to be significantly superior to control group youngsters on a number of measures of social and cognitive development. Fantasy play training was significantly related to a higher incidence of spontaneous sociodramatic play and to higher scores on selected subtests of standard IQ tests, and it facilitated performance on Borke's Revised Interpersonal Perception Test. It also facilitated performance on tasks designed to measure story-sequence memory skills and story verbalization skills. On the other hand, fantasy play did not significantly enhance ability to recall pictures as opposed to objects. In conclusion, it was noted that fantasy play training is a promising and practical intervention method enjoyed greatly by both the children and the adult interventionists. Further use and study of this technique is encouraged.

Recently there has been a resurgence of interest in evaluative intervention research on the use of symbolic or imaginative play as a vehicle to foster the cognitive growth of preschoolers.

Various writers who have attempted to define symbolic or imaginative play, including Feitelson and Ross (1973), Gilmore (1966), Klinger (1969), Millar (1968), Piaget (1962), Singer (1973), and Smilansky (1968), have generally distinguished this form of play from exercise play, mastery play, and games with rules. Although symbolic play can take various forms, it has in common a make-believe or pretend quality; that is, the child uses his imagination to amplify and modify the concrete and immediately present stimulus, for example, the child pretends to wake up a sleeping blanket. Major developmental theorists, for example, Piaget (1962), Vygotski (1967), and Werner (1948), have viewed this type of play as an indispensable step in cognitive development through which the child becomes liberated from stimulus boundedness and thereby advances toward operational and abstract levels of thought.

There have been a number of training studies that have attempted to increase different types of imaginative play behavior in young children in general and especially in preschoolers from deprived social backgrounds (e.g., Feitelson & Ross, 1973; Freyberg, 1973; Marshall & Hahn, 1967; Smilansky, 1968). The rationale for these studies has been based on the assumptions that imaginative play is (*a*) trainable, (*b*) important for the development of other intellectual and social abilities, and (*c*) often undeveloped in young children. In regard to this third point, Smilansky (1968)

[1] The present research is part of a program being supported by a grant from the Spencer Foundation. We wish to express our gratitude to the supervisors of Franklin Preschool, Marlyn Levin and Carolyn Dunlap, and to the teachers, Claudia Rose, Marion Painter, and Betty Keown, for their cooperation throughout the duration of the study.

[2] Requests for reprints should be sent to Eli Saltz, Center for the Study of Cognitive Processes, Wayne State University, Detroit, Michigan 48202.

has suggested that sociodramatic play is directly taught and encouraged in most middle-class homes but is virtually absent in most socially disadvantaged homes.

Smilansky (1968) showed that socially disadvantaged Israeli children could be trained to increase their sociodramatic play. Feitelson and Ross (1973) and Freyberg (1973) also demonstrated that children from lower social and economic levels could be trained to improve their imaginative play. Weikart, Rogers, Adcock, and McClelland (1970), in their influential book, advocated the use of sociodramatic training in programs designed to improve the cognitive functioning of socially disadvantaged preschoolers.

However, despite this increase in interest in various forms of dramatic play activities, there has been almost no systematic attempt to evaluate the contention that any or all of these types of activities can have a benefical effect on cognitive functioning.

A type of dramatic play somewhat akin to Smilansky's sociodramatic play was introduced and evaluated in the intervention study reported here. This form of play can perhaps best be described as thematic-fantasy play (TFP). This type of play is similar to sociodramatic play in that it involves verbal role enactment in a group. However, in TFP, children dramatize traditional folk tales popular with children, for example, *The Three Billy Goats Gruff, Little Red Riding Hood,* etc. Unlike sociodramatic play, then, TFP employs a structured play theme or story plot. It is hypothesized that by providing children with opportunities to enact story sequences, they will be helped to see that events are interrelated and ordered in time and space. Theoretically, this should promote the development of conceptual schemata or the integration of experiences in preschool children.

Thematic-fantasy play is distinguished from sociodramatic play in yet another way; that is, TFP involves real fantasy. In TFP, children are required to imagine and perform behaviors described to them in story narration which are never actually observed in real life. Thus, TFP demands more than what is usually meant by the imitative behavior so central in sociodramatic play; it demands imagined behavior, that is, a transition from symbolic story form into behavior form. Theoretically, this quality of TFP should contribute to later ability to use symbols and think creatively.

An intervention program to teach TFP to disadvantaged children was introduced at Franklin Preschool, Detroit, Michigan, during the winter and spring of 1972. We wished first of all to check the feasibility of training disadvantaged preschoolers in this type of play behavior. Once it became obvious that the children were receptive to this kind of intervention, our goals broadened to include an evaluation of the effects of TFP on various measures of social and cognitive development. Two experimental groups and two control groups were established and training proceeded over a four-month period. Utilizing a 2 × 2 factorial research design with TFP training and dimensionality (D) training as factors, four intervention curricula were generated: (*a*) one group of children received TFP training only; (*b*) one group received D training only, for example, learning to label and categorize stimuli along various dimensions; (*c*) one group received both TFP and D training; and (*d*) one group received neither TFP nor D training, instead this group engaged in story listening and other activities unrelated to either TFP or D training. During the intervention training, records were kept of the childrens' behavior both while performing in the research groups and while engaging in spontaneous free play in the nursery school classroom. Following intervention training, the four research groups were tested on several selected standard instruments and especially designed tasks in order to assess the effects of training.

Method

Subjects

Subjects were preschoolers in Franklin Elementary School, Detroit, Michigan. Four classrooms of approximately 20 children each were involved in the project. The subjects were from primarily lower-economic-class backgrounds and

included approximately 30% southern white, 25% black, 25% northern white, and 20% Chicano. Many of the children were from families who recently moved into the city from the south. During the project, 24 subjects of the original sample pool of 80 subjects left the preschool and 19 subjects enrolled in their places. Replacements were approximately evenly distributed across the four research groups (5, 4, 5, and 5 subjects in the thematic-fantasy play group, dimension training group, mixed group, and control group, respectively) and received about the same amount of training (approximately 10 sessions per replacement subject). In the remaining sample of 75 subjects, there were 44 males and 31 females ranging in age from 2 years 10 months to 5 years 6 months with a median age of 3 years 8 months.

In each of the four classrooms, subjects were divided into four research groups of 5 children each; the groups did not differ significantly on age and pretest scores. The four matched research groups were then designated by chance for thematic-fantasy play (TFP) training; dimensionality training (D); mixed TFP and D training; and control (C) group activities. Each group met as regularly as possible three times a week for 15-minute sessions over a four-month period. All groups had approximately equal exposure to the same three-member team of intervention teachers, any two of which were normally present during a group session.

Training Conditions

A 2 × 2 factorial research design was used in this study. One factor was thematic-fantasy play, the other factor was dimensionality training. This produced the following four groups: (*a*) TFP + D ($n = 19$); (*b*) TFP ($n = 19$); (*c*) D ($n = 19$); and (*d*) C ($n = 18$).

Fantasy (TFP groups). Fantasy subjects were exposed to a TFP curriculum which consisted of systematic training in role enactment of action-type fairy tales, *The Three Pigs, Hansel and Gretel,* etc. Fantasy subjects heard a story read to them, were assigned parts, and enacted the story with intervention teachers prompting, narrating, and at times taking roles in the story themselves. Few props were used other than chairs and tables which represented such things as houses, trees, bridges, etc., depending on the story. The children would dramatize the same story several times over successive group sessions and would take turns playing the various characters in the story. Following role enactment, children would discuss the story plot during which times emphasis was placed on remembering the story sequence and verbalizing the "reasons" for the events that occurred in the story, for example, "Why did the billy goats cross over the bridge?", "Why did the baby bear start to cry?", etc.

Dimensionality (D) training groups. These subjects received systematic training in labeling and classifying activities. Subjects were taught to identify, describe, and classify social and physical stimuli along several dimensions. They were encouraged to verbalize about objects and to discuss the ways in which things go together in an interactional setting.

In some training sessions, the children learned to recognize various forms and changes of physical stimuli. For example, during one session, the preschoolers explored different kinds of grapes and the products of grapes, for example, red, green, and purple grapes, raisins, and grape drinks. In other training sessions, the children learned to recognize various forms and changes of social stimuli. For instance, one meeting involved discussing sex- and age-appropriate clothing and objects. The children matched clothing and objects to cut-out cardboard representations of grandfather, father, mother, little boy, little girl, and baby. In short, in D training, the children were given repeated opportunities to discuss social and physical objects in a group setting.

Fantasy plus dimensionality (TFP + D) groups. These groups received TFP training on 50% of the sessions and D training on 50% of the sessions.

Control (C) groups. The control groups were primarily engaged in listening to the stories used for role enactment by the TFP groups without dramatizing these stories themselves. These subjects also participated in other types of activities unrelated to TFP such as playing with cut-outs, cut-and-paste activities, singing, and so forth.

Evaluation Procedure

Pretests. Before the start of the intervention training, the children were administered (*a*) the Picture Completion subtest of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) and the Visual Reception and Visual Association subtests of the Illinois Test of Psycholinguistic Abilities (ITPA) as rough indicators of nonverbal intelligence and (*b*) the Similarities subtest of the WPPSI as a rough indicator of verbal intelligence.

Postmeasures. Evaluation of the effects of the intervention training utilized play observations and both standardized and specially designed tests. The following is a description of the assessment methods used to evaluate the effects of the intervention program.

1. *Play observations.* On 20 different days during the intervention training, the four classrooms were each observed for about 20 minutes. The observer on these occasions watched for the presence of either sociodramatic or thematic-fantasy play. When such play behavior was spotted, the observer recorded the names of the children involved. A comparison between the proportion of fantasy subjects and control subjects who were observed at least once engaging in such play during the first 10 observations and during the second 10 observations and over all 20 observations was made using chi-square analysis. In addition, the change in the number of preschoolers observed participating in

dramatic free play from the first set of 10 observations to the second set of 10 observations was examined for both fantasy subjects and control subjects separately.

2. *Intelligence subtest postmeasures.* The Similarities and the Picture Completion subtests from the WPPSI and the Visual Reception and Visual Association subtests from the original ITPA were chosen for use in this study. Raw score totals were computed for each subject and this unit was used in analysis with pretest score as a covariate. A randomly selected subset of 34 subjects was administered the intelligence subtests.

3. *Interpersonal Perception Test (IPT).* The IPT which was designed and described by Borke (1971) and then later revised (Borke, 1973) is a test for empathy in young children. This test requires the child to choose the "face" depicting the appropriate affect another child would feel under certain prescribed conditions related to the child in a story form. The four possible selections the child has are "happy," "sad," "afraid," and "mad" faces from which the child selects the most appropriate face. On each of the 23 items of the revised version, we added a "neutral" face possibility and administered this modified version of Borke's IPT to a randomly selected subset of 47 subjects of the present study. Total right in the 23-item test was the score used in statistical analysis.

4. *Picture versus object memory.* Each pair consisted either of two toys or of two pictures of toys. The task involved five paired associates. This task, desgned to measure preschoolers' ability to represent concrete materials versus pictorial representations in memory, was administered to 72 of the 75 subjects involved in the program. Half of the subjects in each of the four research groups, individually tested, received one of two sets of test materials. In Set A, three of the five pairs were object pairs and two were picture pairs; in Set B, three were picture pairs and two were object pairs. The picture pairs in Set A were the object pairs in Set B and vice versa. All subjects then received both object and picture pairs. Number of errors (i.e., failure to respond or incorrect response) for picture pairs and for object pairs over three trials was computed for each subject. Preliminary analysis showed that there were no significant differences on performance on Set A versus Set B, and these two sets were combined within each subject group for additional analysis.

5. *Story memory task.* This specially constructed test was used to assess preschool childrens' ability to remember a story sequence. Subjects, individually tested, were first shown a series of six pictures and were told a story that the pictures depicted. The pictures were then shuffled and the subject was instructed to put the pictures back into the original order. The subject's score was the degree to which his order corresponded to the correct sequence. This score was computed as the tau coefficient translated into Z scores on which statitistical analysis was performed. Forty-four randomly selected subjects were evaluated on this task.

6. *Story telling task.* This also was a specially designed task which was used to evaluate preschoolers' story-telling skill. The subjects, individually tested, were asked to tell three stories from three different series of five pictures each. After each narration, subjects were asked two questions pertaining to the story. These questions were intended to clarify the preschooler's understanding for the reasons behind the actions depicted in the story pictures. The subjects' performance on this task was evaluated in several ways: Computations of total verbal output on the three stories, total number of connectives used, and number of inferences expressed, either spontaneously or in response to questions, were made for each subject. In addition, on Stories 1 and 2, each subject was scored for expressed continuity between pictures on a scale from 0 to 4. Between each consecutive pair of pictures in a story, the subject was credited with 1 point if he stated continuity and 0 if he failed to express continuity. A randomly selected subset of 49 subjects was administered the story-telling task.

Results

The results of the present intervention study are organized in seven parts: (*a*) observations of each of the groups behavior during TFP training; (*b*) observations of spontaneous dramatic play; (*c*) intelligence tests scores; (*d*) Interpersonal Perception Test (IPT) scores; (*e*) picture versus object memory scores; (*f*) story memory scores; and (*g*) story-telling scores.

Group Observations

Since TFP is a relatively unevaluated approach to intervention training, perhaps a few comments about TFP itself are in order.

It was observed that at first most preschoolers found TFP enjoyable but difficult. It was necessary to simplify the stories and participate with the children in the dramatization. It was also found helpful for one of the intervention teachers to narrate the story while the preschoolers role-enacted and to provide the children with lines they had forgotten. Even with extensive coaching and prompting, it was observed that most of the role enactment was nonverbal since, at first, the children proved more willing to act than speak. Moreover, the children showed little appreciation for the story sequence. Often the group remembered only the final scene or the most exciting event in the story. The children had great enthusiasm for action but little idea of why the action

was taking place. In short, during the early stages of TFP intervention, it was apparent that the children had much room for improvement.

With practice, both the children and the intervention teachers became more skillful in TFP. It was discovered that the children functioned better with a minimum of props; it appeared that the use of realistic props at times distracted the children. Consequently, fewer props were employed in TFP. Also, we found that the children seemed to depend on locations in the story remaining constant. It was helpful to identify *places in the playing room* for the children. Knowing locations enabled the preschoolers to orient themselves for role enactment. They seemed to depend on places remaining the same from session to session, although they did not seem to mind switching roles. They even played opposite-sex roles with enthusiasm.

Sometimes the children reenacted events not in the way they happened in the story but as they would have liked them to happen. Amusing instances of this behavior occurred periodically. For example, on one occasion a little girl performed her witch's role almost perfectly until the time came to be pushed in the oven. Suddenly the little actress announced that she was a "good" witch and invited Hansel and Gretel's mother over for coffee and cake! This tendency to assimilate the story, although benevolently accepted when it occurred, was something that was discouraged over the course of TFP training. With practice, the children became more adept at following the sequence of a story and more efficient in TFP in general.

Play Observations

The results of the play observations over 20 sessions indicate that thematic-fantasy play has a significant and positive effect on the preschool child's likelihood of being observed participating in dramatic free play in nursery school. Significantly more fantasy subjects than control subjects were observed at least once during 20 observations engaging in dramatic free play. While 94.7% or 38 of the fantasy subjects were observed in dramatic free play, only 60.5% or 26 of the control subjects were observed. Thus, 5.3% or 2 of the fantasy subjects and 39.5% or 17 of the control subjects were never observed in dramatic free play ($\chi^2 = 14.001$, $df = 1$, $p = .001$). Training in role enactment of fairy tale stories apparently enhances the probability that preschoolers will engage in dramatic social free play in school.

Examination of the changes in frequency of spontaneous TFP over the year for the fantasy subjects and the control subjects indicates that the differences observed in Table 2 are a consequence of the intervention conditions. For example, over the first 10 play observations, 65% of the fantasy subjects were observed in spontaneous TFP. During the second 10 observations, 92% of the fantasy subjects were observed in such activities. Of the 40 subjects present during all 20 observations, 12 subjects participated in TFP during the second 10 observations who had not engaged in TFP during the first 10 observations. A chi-square test for change in frequencies of occurrence (McNemar, 1959) indicated that this increase was significant ($\chi^2 = 7.69, p < .01$).

For the control subjects, on the other hand, the corresponding percentages of subjects observed in spontaneous TFP during the first and second 10 observations were 45% and 40%, respectively; in short, the control subjects showed a small but nonsignificant *drop* in spontaneous TFP.

It may be added that teacher reports substantiate this finding. On several occasions, teachers brought it to the attention of the experimenters that certain children were engaging in sustained dramatic play behavior not only during free play but at other times during the school day. For example, on one occasion an entire snack table, without adult prompting, dramatized a birthday party involving imaginary animals. Inspection showed that 6 of the 8 children so playing were fantasy subjects.

Intelligence Test Scores

The Picture Completion and Similarities subtests of the WPPSI and the Visual Reception and Visual Association subtests

of the ITPA were administered in order to obtain a rough indication of the effects of training on intellectual performance. The sum of the raw scores on these four subtests constituted each child's score. Since testing time was limited, only 34 subjects were evaluated, fewer than on any other measure. To ensure comparability, postmeasures were adjusted for pretest intelligence scores in this analysis. (It might be noted that this was not necessary for the other postmeasures since the four groups had been matched on pretest intelligence scores.) The mean intelligence scores adjusted for pretest scores for the two fantasy conditions were 40.8 and 34.0, respectively, for the TFP children who received dimensions training and those that did not receive such training. For the two conditions that were not involved in the fantasy program, the corresponding means were 37.4 and 28.8, respectively. Analysis of covariance (pretest intelligence of child as covariate) showed that the main effects for TFP and D were significant ($F = 7.35, df = 1/29, p < .011$ and $F = 11.8, df = 1/29, p < .002$, respectively). The interaction was not significant. Both TFP and D training, in other words, appear to have facilitated intellectual functioning. This finding, however, must be interpreted cautiously since a complete intelligence test battery was not administered and relatively few subjects were tested.

Interpersonal Perception Test (IPT)

The IPT is designed to measure the ability of young children to cognitively represent another's affective experience. The present study suggests that TFP training significantly increases the ability of preschoolers to respond correctly on the IPT. The mean scores for fantasy subjects and control subjects were 13.48 and 10.83, respectively ($F = 6.319, df = 1/47, p < .05$). Thus, learning to role-enact characters from children's folk tales apparently increases the ability of preschool children to understand and identify the affective states of other children. Evidently, role enactment training can facilitate role-taking ability.

In contrast to the effects of fantasy training, dimensionality training appears to be unrelated to performance on the IPT. Neither the main effects of dimensionality training nor the Dimensionality × Fantasy interaction approached significance.

Picture versus object memory

The results of the present study replicated previous findings that memory for objects is superior to memory for pictures of these same objects. The relevant F of 76.55 ($df = 1/72, p < .001$) was based on a within-subjects analysis. On the other hand, none of the main effects or interactions involving either fantasy training or dimensionality training approached significance. Apparently, TFP does not improve representational function for this type of rote memory.

Memory for stories

The ability of preschool children to remember a story sequence was assessed using a specially designed story-memory task; the children were required to arrange pictures so as to match an order shown to them earlier. (Note that the initial presentation was accompanied by narrative which explained the sequence of events in the pictures.) The sequences produced by the children were correlated with the original experimental sequence using the tau coefficient. The taus were converted to Z scores so that each child's tau could be considered a score and a mean tau could be computed for each group (recall that untransformed distributions of taus tend to be extremely skewed); next, to facilitate computation, 1.0 was added to each tau (to convert all scores to positive numbers) and each score was multiplied by 100.

The mean score for fantasy subjects was 26.2; for control subjects, 14.9. Analysis of variance indicated that the effects of fantasy training were significant ($F = 4.901, df = 1/44, p < .05$). Neither dimension training nor the Fantasy × Dimension interaction approached significance.

It should be noted that while the difference between the TFP and control groups proved significant on this task, all the preschoolers found the task extremely difficult.

Story-telling task

Young children are known to view the events of their lives as a series of "still"

pictures, with little sense of causal relationship between the events. The primary purpose of the present task was to determine if TFP helped children see the relationship between events. A secondary purpose was to determine if TFP affected some of the gross indices of verbal ability. The task consisted of asking the children to tell the simple stories illustrated in a sequence of cartoon pictures.

The first measure of ability to see causal relations was obtained by scoring Stories 1 and 2 for consistency of theme from picture to picture; for example, did the story to Card 2 follow from the story to Card 1 in a sequence of pictures. The children in the TFP groups performed better than the children in the control groups on both stories. On Story 1, the TFP children obtained a mean score of 1.79, which indicates continuity on about 45% of their transitions from one picture to the next; the control children obtained a mean of .89, which indicates continuity on 22% of their transitions. In Story 2, the respective means were 2.57 (66% of the transitions) and 2.38 (59% of the transitions). The difference for Story 1 was significant ($F = 5.14$, $df = 1/49$, $p < .028$); the difference for Story 2 did not approach significance. None of the other effects (viz., main effects for dimensions nor the Dimensions × Fantasy interaction) approached significance for either story.

The use of inference was a much more stringent measure and required that a child overtly connect an event in one picture to some other event in a different picture in such a way that it was clear that the two pictures were part of the same story. The scores on this measure proved to be extremely skewed, requiring analysis by means of a nonparametric technique. Analysis of the stories showed that 95.5% of the 22 fantasy children made at least one such inferential statement in the two stories combined; only 63.7% of the 22 control children used such statements. This difference yields a chi-square of 5.028 ($p < .05$) corrected for continuity.

The use of connectives was also seen as indicating an attempt to integrate the events of a story, and so the frequency of connectives was analyzed. These scores also proved to be extremely skewed, necessitating the use of a nonparametric test of significance. With 22 children per group, 100% of the fantasy group used connectives at least once while only 67.7% of the controls used them. The resulting chi-square was corrected for continuity ($\chi^2 = 7.486$, $p < .01$).

Finally, turning to total verbal output, the mean number of words used in story telling by the TFP children and the control children were 71.7 and 49.6, respectively, over the two stories. This difference was highly significant ($F = 7.925$, $df = 1/40$, $p < .01$). The effect was approximately the same for each of the two stories, with no interaction between stories and conditions.

In conclusion, the data for the story-telling task suggest that the fantasy children made more of an attempt to connect and integrate the events in telling a story. In support of this conclusion it should be mentioned that after each child told his story, he was asked questions about it. It was found that the children in the fantasy group appeared to possess greater comprehension than the control subjects for the reasons behind the actions depicted in the series of pictures.

Summary and Conclusion

This study was conducted primarily to utilize and evaluate thematic-fantasy play and dimensions training as an intervention technique for socially disadvantaged preschoolers. Our preliminary observations and test results indicate that thematic-fantasy play distinguishes itself as an enjoyably workable and promising intervention tool which significantly affects a number of positive behaviors and abilities in preschool children. Thematic-fantasy play, unlike dimensionality training, was found to be significantly associated with a higher incidence of spontaneous social-dramatic play, superior performance on Borke's (1973) Interpersonal Perception Test, and better story-memory and story-telling skill on specially constructed tasks.

The effects of both dimensionality training and fantasy play training on intelligence were more equivocal. Although a relatively large difference in obtained scores occurred

between fantasy, dimension, and control groups and although these effects were statistically significant, it should be noted that sample size was small and that only selected subtests were employed rather than a full scale test. More work is needed on this issue.

Finally, there was no indication that fantasy training or dimensionality training influenced ability to recall pictures as opposed to objects. If the inferior memory for pictures, as opposed to objects, is due to poorly developed ability for mental representation, the type of representation required appears to be different from that involved in either fantasy play or dimensionality training.

The present report summarized the findings from the first year of a long-term research project. While our conclusions must be tentative at this time, the technique of fantasy training appears to be more promising than the techniques of dimensionality training. Any benefits specific to dimensionality training are not indicated by the outcome measures of interest in this study.

Preschool intervention programs in the United States and abroad have utilized a rather broad range of techniques and methods intended to compensate for cognitive and socioemotional deficits incurred by conditions of poverty. One of the major assets of thematic-fantasy play as an intervention technique is its appeal to preschool-age children. Our observations indicated that almost all of the children found listening to, discussing, and then dramatizing "action-filled" fairy tales very rewarding. The children appeared to regard thematic-fantasy activities as fun not work. The children's enthusiasm was shared by the intervention teachers, who enjoyed playing with the children, with the net effect that thematic-fantasy play proved to be a very encouraging, workable, and promising intervention technique, one that deserves further use and study.

REFERENCES

Borke, H. Interpersonal perception of young children: Egocentrism or empathy? *Developmental Psychology*, 1971, **5**, 263–269.

Borke, H. The development of empathy in Chinese and American children between three and six years of age: A cross-cultural study. *Developmental Psychology*, 1973, **9**, 102–108.

Feitelson, D., and Ross, G. The neglected factor—play. *Human Development*, 1973, **16**, 202–223.

Freyberg, J. Increasing the imaginative play of urban disadvantaged kindergarten children through systematic training. In J. L. Singer (Ed.), *The child's world of make believe: Experimental studies of imaginative play*. New York: Academic Press, 1973.

Gilmore, J. Play: A special behavior. In R. N. Haber (Ed.), *Current research in motivation*. New York: Holt, 1966.

Klinger, E. Development of imaginative behavior: Implications of play for a theory of fantasy. *Psychological Bulletin*, 1969, **72**, 277–298.

Marshall, H., & Hahn, S. Experimental modification of dramatic play. *Journal of Personality and Social Psychology*, 1967, **5**, 119–122.

McNemar, Q. *Psychological statistics*. (2nd ed.) New York: Wiley, 1959.

Millar, S. *The psychology of play*. Harmondsworth, England: Penguin, 1968.

Piaget, J. *Play, dreams, and imitation in children*. New York: Norton, 1962.

Singer, J. *The child's world of make-believe: Experimental studies of imaginative play*. New York: Academic Press, 1973.

Smilansky, S. *The effects of sociodramatic play on disadvantaged preschool children*. New York: Wiley, 1968.

Vygotski, L. Play and its role in the mental development of the child. *Soviet Psychology*, 1967, **5**, 6–18.

Weikart, D., Rogers, L., Adcock, C., & McClelland, D. *The cognitively oriented curriculum: A framework for preschool teachers*. Urbana, Ill.: University of Illinois, 1970. (ERIC-NAEYC Publication in Early Childhood Education)

Werner, H. *Comparative psychology of mental development*. New York: Science Editions, 1948.

(Received November 2, 1973)

39 years of continuous service

for Discriminating Test Users

Edited by OSCAR K. BUROS

NEW Tests in Print II.

In press — to be published late 1974 — a comprehensive index to tests, test reviews, and the literature on specific tests; about 1,000 two-column pages; $50.00. TIP II includes a bibliography of approximately 2500 tests published in English as of March 1974; over 16,000 new references through 1971 on the construction, use, and validity of specific tests; a cumulative name index for each test with references; a complete title index of all published tests (including out of print); and a name index of all authors of tests, reviews, excerpts, and references in the *Mental Measurements Yearbook* series.

1972 The Seventh Mental Measurements Yearbook

1972; 2032 pages in 2 volumes; $70 per set. "Indispensable . . . comprises all the values of its predecessors, but in larger measure . . . simple, clear indexing system . . . statistics can give only the most inadequate picture . . . the newest edition lists 1,157 tests with 12,372 references. For 546 of the tests, there are 798 original reviews by 439 reviewers, supplemented by another 181 reviews excerpted from 39 journals." — *Test Service Bulletin of The Psychological Corporation,* May 1972.

1970 Personality Tests and Reviews

1970; 1690 pages; $45.00. "The second of four monographs covering specific sections of the first six *Mental Measurements Yearbooks.* * A great deal of new material is also presented for the first time." — Wayne H. Holtzman, *Am J Orthopsychiatry,* July 1970.

1968 Reading Tests and Reviews

1968; 542 pages; $20.00. "Brings together in convenient form the enormous amount of bibliographical material about reading tests of the past and present that have been available heretofore only by consulting separately the six *Mental Measurements Yearbooks.*" — Frederick B. Davis, *Ed & Psychol Meas,* Autumn 1969.

Earlier Volumes Needed for Full Coverage

1965 **The Sixth Mental Measurements Yearbook.** Reprinted 1971; 1751 pages; $45.00.
1961 **Tests in Print.** Reprinted 1974; 508 pages; $15.00.
1959 **The Fifth Mental Measurements Yearbook.** 1321 pages; $35.00.
1953 **The Fourth Mental Measurements Yearbook.** Reprinted 1974; 1188 pages; $30.00.
1949 **The Third Mental Measurements Yearbook.** Reprinted 1974; 1062 pages; $25.00.
1941 **The 1940 Mental Measurements Yearbook.** Reissued 1972; 672 pages; $22.50.
1938 **The 1938 Mental Measurements Yearbook.** Reissued 1972; 430 pages; $20.00.

Order directly from
THE GRYPHON PRESS 220 Montgomery Street, Highland Park, N. J. 08904

CHARLES C THOMAS · PUBLISHER

CLASSROOM DISCIPLINE: A Positive Approach *by* **Susan Bray Stainback,** *Hope Haven Children's Hospital, Jacksonville, and* **William Clarence Stainback,** *Univ. of North Florida, Jacksonville.* The primary focus of this book is on straightforward, specific, workable techniques for classroom control. The authors outline exactly what has to be done in the real, live classroom situation to control disruptive behavior. Viable alternatives to outmoded techniques of classroom control, such as threats and corporal punishment, are also provided. *'74, 180 pp., $5.50, paper*

VOCATIONAL REHABILITATION OF THE MENTALLY RETARDED: A Book of Readings. *Edited by* **Lloyd Keith Daniels.** *Central Connecticut State College, New Britain. (104 Contributors)* Included in this work are a wide variety of topics essential to broadening knowledge regarding the retarded. The articles were chosen for their logical and consistent arrangement of ideas and each selection either describes a program or presents a systematic procedure for resolving an issue or solving a problem of traditional concern to personnel engaged in the vocational rehabilitation of the mentally retarded. *'74, 648 pp. (7 x 10), 11 il., 60 tables, $22.50*

ORGANIZATION AND ADMINISTRATION OF EDUCATIONAL PROGRAMS FOR EXCEPTIONAL CHILDREN *by* **B. R. Gearheart,** *Univ. of Northern Colorado, Greeley.* Written for use as a basic college text for those training to become supervisors or administrators of special education programs, this volume should be of interest and value to all those who come in contact with exceptional children. The review of federal legislation for the handicapped and its relationship to the evolving role of federal leadership, special financing for programs, and the scope and function of both state and local special education administration add to the practical applicability of this volume. *'74, 272 pp., 1 table, $11.75*

TRAINING RETARDED BABIES AND PRESCHOOLERS *by* **Thomas F. Linde,** *V. A. Hospital, Knoxville, Iowa, and* **Thusnelda Kopp,** *Washoe County Association for Retarded Children, Reno, Nevada.* Based on the premise that most mentally retarded babies and preschoolers possess important potential to make productive contributions to their families, neighborhoods and communities, this volume is designed for public health personnel, psychologists, parents and others initiating and implementing activities for shaping basic perceptual, motor and social skills. *'74, 200 pp., 12 il., $12.75*

EFFECTIVE TEACHING METHODS FOR AUTISTIC CHILDREN *by* **Rosalind C. Oppenheim,** *The Rimland School for Autistic Children, Evanston, Illinois. Foreword by* **Bernard Rimland.** This reference book for parents, special educators and professionals in related disciplines emphasizes intruding on the child's withdrawal and overcoming resistance and negativism; establishing control, such as dealing with tantrums and disruptive behavior; the teacher-child relationship in the learning situation; and methods for teaching reading, writing and arithmetic. *'74, about 104 pp.*

THE RIGHT-TO-EDUCATION CHILD: A Curriculum for the Severely and Profoundly Mentally Retarded *by* **Donald G. Myers,** *Marywood College, Scranton, Pennsylvania, and* **Michael E. Sinco** *and* **Ellen Sommerton Stalma,** *both of Northeastern Intermediate Unit Nineteen, Scranton, Pennsylvania. Foreword by* **P. M. Mensky.** *(2 Contributors)* For teachers, college students, parents and those interested in working with the severely and profoundly mentally retarded, this book offers techniques based on well-established principles of learning. Step-by-step, carefully detailed procedures are provided for helping these children achieve better sensory development, motor development, self-care development and language readiness development. *'74, 248 pp., 41 il., $9.75*

Orders with remittance sent, on approval, postpaid

301-327 East Lawrence Avenue • Springfield • Illinois • 62717

Library
26/2/75

Journal of Educational Psychology

VOLUME 66 NUMBER 5

OCTOBER 1974

AMERICAN PSYCHOLOGICAL ASSOCIATION

EDITOR
JOANNA WILLIAMS, *Teachers College, Columbia University*

ASSOCIATE EDITOR
SAMUEL BALL, *Educational Testing Service*

ADVISORY EDITORS
PAUL R. AMMON, *University of California, Berkeley*
JERE BROPHY, *University of Texas at Austin*
C. VICTOR BUNDERSON, *Brigham Young University*
LEONARD S. CAHEN, *Educational Testing Service and Far West Laboratory for Educational Research and Development*
LOIS-ELLIN DATTA, *National Institute of Education*
FREDERICK B. DAVIS, *University of Pennsylvania*
FRANCIS J. DIVESTA, *Pennsylvania State University*
ROBERT EBEL, *Michigan State University*
DAVID ELKIND, *University of Rochester*
LAWRENCE T. FRASE, *Bell Laboratories*
N. L. GAGE, *Stanford University*
ROBERT GAGNÉ, *Florida State University*
ROBERT GLASER, *University of Pittsburgh*
MIRIAM L. GOLDBERG, *Teachers College, Columbia University*
EDMUND W. GORDON, *Teachers College, Columbia University*
VERNON C. HALL, *Syracuse University*
ARTHUR R. JENSEN, *University of California, Berkeley*
PHYLLIS A. KATZ, *City University of New York, Graduate School and University Center*
HERBERT J. KLAUSMEIER, *University of Wisconsin—Madison*
ROBERT E. LANA, *Temple University*
GERALD S. LESSER, *Harvard University*
JOEL R. LEVIN, *University of Wisconsin—Madison*
FREDERICK J. MCDONALD, *Educational Testing Service*
WILBERT J. MCKEACHIE, *University of Michigan*
FRANK B. MURRAY, *University of Delaware*
LAUREN B. RESNICK, *University of Pittsburgh*
WILLIAM D. ROHWER, JR., *University of California, Berkeley*
ROBERT ROSENTHAL, *Harvard University*
ERNST Z. ROTHKOPF, *Bell Laboratories*
S. JAY SAMUELS, *University of Minnesota*
SEYMOUR B. SARASON, *Yale University*
MELVYN I. SEMMEL, *Indiana University*
IRVING SIGEL, *Educational Testing Service*
HARRY SILBERMAN, *University of California, Los Angeles*
RICHARD E. SNOW, *Stanford University*
GEORGE G. THOMPSON, *Ohio State University*

JOURNAL STAFF
HAROLD P. VAN COTT, *Managing Editor*
ANITA DEVIVO, *Executive Editor and General Manager*
LINDA B. BAROODY, *Supervising Technical Editor*
SHERRI RICARDO, *Technical Editor*
LYNN D. SPIER, *Acting Manager, Publications Business Services*
ANNE REDMAN, *Subscription Manager*

Published bimonthly in one volume per year, six issues (February, April, June, August, October, and December), by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D. C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D. C., and additional mailing offices.

The **Journal of Educational Psychology** publishes original investigations and theoretical papers dealing with learning and cognition, especially as they relate to problems of instruction, and with the psychological development, relationships, and adjustment of the individual. Preference is given to studies of the more complex types of behavior, especially in or relating to educational settings. Journal articles pertain to all levels of education and to all age groups.

Manuscripts: Submit manuscripts in triplicate to the Editor, Joanna Williams, P.O. Box 51, Teachers College, Columbia University, New York, New York 10027, according to instructions provided elsewhere in the Journal.

Copyright and Permissions: All rights reserved. Written permission must be obtained from the American Psychological Association for copying or reprinting text of more than 500 words, tables, or figures. Permissions are normally granted contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of $10 per page of APA material. Permission fees are waived for authors who wish to reproduce their own articles, or when material is reproduced in limited quantities for instructional purposes. Address requests for permission to the Permissions Office, American Psychological Association, 1200 Seventeenth Street, N. W., Washington, D.C. 20036.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1974: $24 domestic, $25 foreign, $5 single issue. APA member rate: $10. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive galley proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Department, APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P. O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. Undelivered copies resulting from address changes will not be replaced; second-class mail is not forwarded unless subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Diary No. 867
D
31/3/1
Bureau

JOURNAL OF EDUCATIONAL PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

Volume 66 October 1974 Number 5

Notice to APA Authors 631

Learning by Televised "Plaza Sesamo" in Mexico 632
ROGELIO DIAZ-GUERRERO AND WAYNE H. HOLTZMAN

Acquisition Processes and Resilience under Varying Testing Conditions for Structurally Different Problem-Solving Procedures 644
RICHARD E. MAYER

Longitudinal Changes in Intellectual Functioning in Preschool Children: Schooling- and Age-Related Effects 657
L. R. GOULET, KERRY G. WILLIAMS, AND CARL M. HAY

Doll Preferences: An Index of Racial Attitudes? 663
PHYLLIS A. KATZ AND SUE ROSENBERG ZALK

Indirect Review and Priming Through Questions 669
ERNST Z. ROTHKOPF AND MARJORIE J. BILLINGTON

Sex Differences in Comprehension of High- and Low-Interest Reading Material 680
STEVEN R. ASHER AND RICHARD A. MARKELL

Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies 688
DONALD B. RUBIN

Relationship of Three Cognitive Style Dimensions to Elementary Reading Abilities 702
DOUGLAS R. DENNEY

Effects of Cognitive Intervention on Latent Ability Measured from Analogy Items 710
SUSAN E. WHITELY AND RENÉ V. DAWIS

Auditory–Visual Integration and Reading Performance in Lower-Social-Class Children 718
GERALD W. JORGENSON AND ELIZABETH M. HYDE

Impact of Coeducation on "Fear of Success" Imagery Expressed by Male and Female High School Students 726
RONALD WINCHEL, DIANE FENNER, AND PHILLIP SHAVER

Relationships among Perceptual-Motor Tasks: Tracing and Copying 731
MARCIA BERNBAUM, JACQUELINE GOODNOW, AND ELYSE LEHMAN

Acquisition and Nonspecific Transfer Effects in Prose Learning as a Function of Question Form 736
JAMES F. VOSS

Student Teams and Learning Games: Their Effects on Cross-Race and Cross-Sex Interaction 741
DAVID L. DEVRIES AND KEITH J. EDWARDS

Effect of Intent-Action Discrepancy and Student Performance Feedback on Teacher Behavior Change 750
WALTER DOYLE AND JUDITH MCNALLY REDWINE

(*Continued on next page*)

(*Continued from previous page*)

Effects of a Four-Year Follow Through Program on Economically Disadvantaged Children 756
WILLA D. ABELSON, EDWARD ZIGLER, AND CHERYL L. DEBLASI

Abstractive Processes in the Remembering of Prose 772
RONALD E. JOHNSON

Encoding Speed, Rehearsal, and Probed Recall of Dyslexic Boys 780
CARL SPRING AND CAROLYN CAPPS

Effects of Learning Prose with Part versus Whole Presentations of Instructional Objectives 787
ROBERT KAPLAN

Cognitive Style as a Variable in School Learning 793
JACK E. ROBINSON AND JERRY L. GRAY

List of Manuscripts Accepted 709

Announcement 800

SUGGESTIONS TO CONTRIBUTORS

In preparing articles for publication in the Journal, authors are strongly advised to follow the general directions given in the *Publication Manual of the American Psychological Association* (2nd ed.). All articles must be preceded by an abstract of 100–120 words. Instructions for preparation of the abstract appear in the Manual on p. 15. Special attention should also be given to the preparation of references (pp. 58–67, 122–130). In reference lists, give journal titles in full; do not abbreviate. *All copy must be double-spaced, including the references.* All manuscripts should be submitted in *triplicate*, one of which should be an original typed copy. Since the reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, the date the manuscript is submitted, and footnotes identifying the author or his affiliation. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Every effort should be made by the author to see that the manuscript itself contains no clues as to his identity. Dittoed and mimeographed copies are not acceptable and will not be considered. Original figures should be high-contrast, glossy, black and white prints of professional quality. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail and to check carefully the typing of the final copy.

Notice to APA Authors

The American Psychological Association announces publication of the second edition of the *Publication Manual* in August 1974. This new edition, which supersedes the 1967 Manual, will be adopted by the 14 APA journals in 1975.

The new Manual is more comprehensive than the previous edition. It updates APA publication policies and procedures and incorporates changes in editorial practice since 1967. For instance, APA now sends many authors their copy-edited manuscripts for review before they are set into type, and some editors now use blind review procedures. The second edition also includes up-to-date statements of the coverage of each APA journal including the *Journal of Experimental Psychology* which will be published in four separate sections in 1975.

The new *Publication Manual* initiates several changes in APA style. These changes are announced in the August 1974 *American Psychologist* and will be introduced in the APA journals in January 1975. During the period of transition to the new style, authors should note that (*a*) all manuscripts published in 1974 will be copy-edited according to the 1967 Manual, (*b*) manuscripts accepted in 1974 and published in 1975 will be copy-edited to conform to the new Manual. Starting in 1975, accepted manuscripts that depart significantly from the Manual will be returned to authors for correction.

Authors will be encouraged by the changes in the second edition. The new APA style simplifies reference forms; eliminates unnecessary underlines, brackets, and other devices; supports appropriate use of "I" and "we"; and generally clarifies typing requirements. Material is arranged for maximum convenience to authors and typists, and all sections are cross-referenced and indexed.

The new *Publication Manual* is available after August 1 for $3. Send orders to APA Publication Sales, 1200 Seventeenth Street, N.W., Washington, D.C. 20036. Orders of $15 or less must include payment unless they are submitted on institutional purchase order forms.

Journal of Educational Psychology
1974, Vol. 66, No. 5, 632-643

LEARNING BY TELEVISED "PLAZA SESAMO" IN MEXICO[1]

ROGELIO DIAZ-GUERRERO[2]

Centro de Investigaciones Psicopedagogicas, A. C. Mexico City, Mexico

WAYNE H. HOLTZMAN

Hogg Foundation for Mental Health, University of Texas

A controlled experimental study was made of the effects of "Plaza Sesamo," a Spanish version of "Sesame Street," during its first telecast season in Mexico City. A total of 221 children from three different lowest-class day-care centers, equally divided by ages 3, 4, and 5, and by sex, were randomly assigned to experimental and control groups. Complete data were later obtained for 173 of these children. A battery of nine tests was individually administered pre-, during, and posttelecast. Measures of attention to the program and of attendance were also taken. Highly significant differences were found for specific achievement tests dealing with general knowledge, numbers, letters, and words as taught by "Plaza Sesamo." Significant differences were also found for five cognitive tests only indirectly related to "Plaza Sesamo" as well as for Oral Comprehension, a test completely independent of the program content. The largest differences occurred in 4-year-olds; the smallest in 3-year olds. The rate of learning was consistently faster for the experimental groups than the controls across the three testing periods. Amount of attention correlated as high as .49 with gains as measured by several tests. Low but significant correlations were found between attendance and amount of gain.

The public acceptance of "Sesame Street" for preschool education by television in 1970 marked the beginning of a new era. Close collaboration of educators, psychologists, sociologists, psychiatrists, actors, artists, producers, and other specialists resulted in large-scale efforts to design, produce, evaluate, and broadcast children's programs aimed at preparing millions of children to cope more adequately with early school learning. The initial evaluations of "Sesame Street" by Ball and Bogatz (1970; Bogatz & Ball, 1971) clearly demonstrated significant gains associated with viewing "Sesame Street," although the precise meaning of these gains has been the subject of controversy (Cook, 1972). Some criticisms of the initial evaluations of "Sesame Street" arise from the weaknesses inherent in any large-scale field study where certain variables cannot be rigorously controlled.

The original "Sesame Street" program has been translated into many languages with little alteration of the visual images in the film itself. The development of "Plaza Sesamo" as a completely new production adapted to Mexican culture has presented a unique opportunity for evaluative research using a true experimental design, thereby overcoming some of the difficulties in the early studies of "Sesame Street." These formative evaluations serve also as a prelude to more extensive summative evaluations to be undertaken on a large-scale basis in the near future.

The present experiment, using preschool children in day-care centers, was designed by the Centro de Investigaciones Psicopedagogicas, Asociacion Civil (CIPAC), Mexico City, Mexico. Experimental and control groups were established in day-care centers by random selection of children. In addition,

[1] This research was supported in part by the Ford Foundation Grant 730-0206. We wish to acknowledge the continued cooperation of Lic. Juan M. Mendoza Chavez and Arq. Guillermo Rossell from the Secretaria de Salubridad y Asistencia, without whom this study could not have been made. We are also deeply grateful to Raul Bianchi and Rosario Ahumada de Diaz for their assistance in implementing the research design, and to Donald Witzke for his assistance in analysis of the data.

[2] Requests for reprints should be sent to R. Diaz-Guerrero, Georgia 123, Mexico 18, D. F.

the experiment provided an opportunity to try out new content-oriented tests designed for "Plaza Sesamo" and independent tests of cognitive ability which were adapted for use with Spanish-speaking preschool children.

Method

Subjects

The directors of eight day-care centers under the auspices of the Secretaria de Salubridad y Asistencia were approached for cooperation in the experiment. Of the four centers expressing a wish to cooperate, three had a sufficient number of children of the right socioeconomic status and proportionality of ages and sexes to fulfill the requirements of the research design. Emphasis was placed upon children from families in the lowest social class. It was especially important to establish the value of children's educational television programs for this largely illiterate but highly important segment of the population.

The mothers of all children in the three day-care centers within the age range of from 3 to 5 were interviewed by research assistants using a short demographic questionnaire. Nearly all families fitted the definition of lowest social class; that is to say, the father (or head of the family) had not completed primary school and was employed as an unskilled worker. At pretest, there were 221 children almost equally divided by age group and sex.

Experimental Design

One-half of the children, by age group and sex, were randomly selected to constitute the experimental group while the other half was designated as a control group. Children in the experimental group were placed in rooms of each day-care center where they viewed "Plaza Sesamo" programs for 50-minute periods, 5 days a week. The entire series of 130 programs took 6 months to complete. While the children in the experimental group were watching "Plaza Sesamo," the children in the control group were viewing cartoons and other noneducational TV programs in a separate room.

The impact of "Plaza Sesamo" upon the children who viewed it was evaluated by a series of individually administered tests given uniformly to children in both the experimental and control groups at three points in time: (*a*) immediately prior to the exposure to "Plaza Sesamo" or the control films; (*b*) seven weeks later; and (*c*) six months later at the end of the experiment. An interval of seven weeks between the first two testing sessions was chosen since the earlier studies of "Sesame Street" used this size interval. Any initial effects should be apparent within this time. In addition, daily ratings were made by trained observers using a 4-point scale for measuring the degree of attention exhibited by the children in the experimental group. Due to limited resources, ratings were not made on every child every day; rather, six children were selected randomly within each of two day-care centers for observation each day. Over the period of 130 programs, every child was rated a number of times in these two day-care centers. One additional variable consisted of the number of absences for each child in both the experimental and control groups.

TABLE 1
DISTRIBUTING OF CHILDREN COMPLETING THE EXPERIMENT

Age/group/sex	Day-care center			
	Las Aguilas	San Angel	Tizapan	Total
3-year-olds				
Experimental				
male	4	5	3	12
female	3	5	4	12
Control				
male	6	5	5	16
female	6	4	2	12
4-year-olds				
Experimental				
male	6	6	2	14
female	7	3	3	13
Control				
male	5	6	4	15
female	7	4	3	14
5-year-olds				
Experimental				
male	6	6	5	17
female	6	7	4	17
Control				
male	8	6	5	19
female	5	4	3	12
Total	69	61	43	173

The number of children remaining in the experiment for the entire six-month period can be seen in Table 1. Of 221 children in the initial sample, 173 completed the experiment. No discernible bias due to dropouts could be discovered.

Individual Test Battery

Most of the tests were translated and adapted from similar tests developed by the Educational Testing Service for use in its evaluation of "Sesame Street." Some had also been used in earlier formative evaluation studies during the development of "Plaza Sesamo" in Mexico. The tests were given by carefully trained preschool teachers and psychological assistants at each of three points in time, the pretest, the during test, and the posttest. A brief description of each follows.

General Knowledge. This test contains 37 items yielding a possible range of scores from 0 to 37. Items consist of questions such as "What do we smell with?", "What do we see with?", "What is

the name of different parts of the body?", and "Which one of these four objects is the heaviest?"

Numbers. This test contains 60 items dealing with such areas as counting, recognizing numbers, naming numbers, and finding a specific stimulus number among four others. Total score on the test can range from 0 to 60.

Relations. This test contains eight items dealing with questions such as "Which bird is closest to the cage?", "Which dog is going across the fence?", and "Which monkeys are between the trees?" Total score ranges from 0 to 8.

Embedded Figures. This preschool version of Herman Witkin's Embedded Figures Test is now commercially available (Coates, 1971). The instructions were translated into Spanish and pilot tested. Each of the 24 items consists of geometric designs in which a simple figure is embedded in a larger design. The test measures the ability of an individual to extricate hidden figures from the context. Total score ranges from 0 to 24.

Parts of the Whole. This test contains 10 items. In each item there is a picture with separated parts and four whole objects (or figures or letters). The child is asked to say which of the whole pictures can be formed by the separated parts when put together. Total score ranges from 0 to 10.

Ability to Sort. This test contains 16 items; the total score can range from 0 to 16. The nature of the test can be illustrated by the following example: A card is presented with four figures, one with two shoes and the other three with only one shoe. The child is then asked which of the four figures is different from the other three.

Letters and Words. Involves the matching of letters as well as the recognition and labeling of letters. In addition, a number of the items deal with word recognition, the reading of words, and the matching of figures with words. A total of 51 items yields a range of scores from 0 to 51.

Classification Skills. This test contains 24 items yielding a range of scores from 0 to 24. A typical item consists of two cards, one of which contains pictures of three objects (fruits, animals, etc.) that define a specific class. The other card contains drawings of four objects, only one of which can be correctly classified with the three on the first card. The child is asked to tell which of the four in the second card belongs with the three in the first.

Oral Comprehension. Developed experimentally by Herschel Manuel at The University of Texas at Austin for use with bilingual children in either Spanish or English (Manuel, 1972), this test contains 35 items, yielding a total score ranging from 0 to 35. A typical item consists of a set of four pictures of objects, animals, etc., about which the examiner asks simple questions. The child's answer reveals whether or not he understands the instructions or can perceptually identify the figure. Unlike a test dealing with perceptual discrimination, the objects in this test are sufficiently simple and different that any young child can easily recognize them.

Viewing Activities

Experimental and control children were treated alike in every sense except for the programs viewed on television beginning at 3 o'clock each afternoon. The experimental groups were taken into a separate room to see "Plaza Sesamo" while the control groups were taken to other classrooms to see cartoons or other noneducational children's programs that were available over television at that hour. The television sets were 25-inch, black-and-white, new receivers placed in the front of the room elevated enough so that every child could see the screen and hear with no difficulty.

Two minor exceptions to the general rule that experimental and control children were treated alike had to be made because of the space limitations within the day-care centers. The experimental groups saw "Plaza Sesamo" in groups averaging in size about 15 children, while the control groups viewed cartoons in groups of approximately twice this size. This type of distribution might favor slightly the experimental cases in any comparisons, although in both experimental and control groups there was no difficulty in viewing the television screen. The second difference between the experimental and control groups arose from the fact that the control cases were kept in the day-care centers until 7 o'clock in the evening while the experimentals were released shortly after 5 o'clock. This difference was necessary to prevent the control children from viewing "Plaza Sesamo" on another channel when it was broadcast each evening from 6 to 7 o'clock. Parents of the control children were informed that their children would be undertaking special activities and games supervised by the assistants and, therefore, that they should not come for their children until 7 P.M. The games and social activities were designed to be as noncognitive as possible.

As a further precaution against uncontrolled factors influencing the experiment, a few weeks after the experiment began, a survey was made of all the parents to determine the extent to which television sets were used in the home. It was found that the incidence of television sets was essentially the same for both groups—67% in the experimentals and 63% in the controls. Still another possibility was that children when absent from the day-care center might be viewing "Plaza Sesamo" at home. The survey revealed that 50% of the experimental absentees viewed "Plaza Sesamo" at home while only 34% of the controls looked at it when absent due to sickness. The average number of absences for the entire sample of children was only 24. But, the fairly large standard deviation (16) indicates that a small number of children were absent fairly frequently, usually because of colds and flu.

Some unexpected events threatened to disrupt the experiment in spite of every precaution. During the six months of the study, there were two official vacations. The first one was the Easter va-

cation lasting only four days. The central authorities and day-care personnel were convinced by the research staff of the necessity to keep the day-care centers open on a regular schedule in spite of the Easter vacation, and parents were encouraged to bring their children to the day-care centers. The second vacation was much more difficult, lasting a period of 11 days in May, toward the end of the experiment. Ordinarily, all of the personnel within the day-care centers would be away for this holiday. There was a real danger that children in the control groups would view "Plaza Sesamo" on their home television sets if they were not kept in the day-care centers during this 11-day period. The authorities were persuaded to keep the day-care center open in the afternoons from 2 to 7 P.M. in spite of the holiday. As an inducement to parents to cooperate in the plan, the personal color television set of the principal investigator was donated as a prize to be raffled off among the parents of those children who had perfect attendance records during the 11-day period. Attendance during these days reached the highest peak of the entire study for both experimental and control groups. The television set was won by the father of one of the control children and he appeared the next day in his best attire, posing happily with his wife and child for photographs and basking in his enviable new status as the proud possessor of a color TV set.

It is believed that the extraordinary steps taken to prevent any contamination of the control groups and to insure continuous viewing of "Plaza Sesamo" by the experimental groups were fully justified. In no case did it appear that the experimental design had been compromised. We believe the results can be interpreted with confidence.

Results

Three different statistical methods were employed in analyzing the data. The first method consisted of a large intercorrelation matrix containing all the test variables at each of the three points in time when measurement took place—pretest, during the treatment, and posttest. In addition, the sex of the child and identification of the day-care center were cast in the form of dichotomous variables and entered into the correlation matrix. The resulting point-biserial correlations were utilized for testing the significance of the difference between experimental and control, between male and female, and between the three day-care centers (one vs. the other two combined), since the results are equal to those obtained by Fisher's t tests. The last variables included in the matrix consisted of the mean attention score for each child in the experimental group (with a reduced N, of course) and the actual number of absences.

The computer program allowed for a variable number of cases for each correlation in the matrix and printed out a separate triangular matrix containing the number of cases for each correlation. The resulting supermatrix consisted of three matrices of intercorrelations of special interest, the intercorrelations for the pretest measures, the intercorrelations for the during-treatment test measures, and the intercorrelations for the posttest measures. The remaining matrices in the supermatrix consisted of correlations across time, such as the correlation of General Knowledge at pretest with Numbers at posttest. The diagonals in these matrices consist of test–retest correlation coefficients for each of the measures employed in the study. Only the highlights of this correlational analysis can be presented in the present report.[3]

For clarity of discussion, the nine tests in the battery given each child can be classified into three different types. The first category can be called content-achievement tests since all of them are criterion measures of skills specifically taught in the "Plaza Sesamo" programs. General Knowledge, Numbers, and Letters and Words were constructed to measure the achievement of each child in each of these areas. The second category of tests involves measures indirectly related to "Plaza Sesamo" and can be called cognitive-content measures. Five tests are classified in this category—Relations, Parts of the Whole, Ability to Sort, Classification Skills, and Embedded Figures. These tests are content-related to "Plaza Sesamo" in the sense that one of the goals of "Plaza Sesamo" is to increase the child's abilities in these areas. But the tests themselves are not specifically criterion measures. The third

[3] Complete results for the correlational studies, as well as other statistical analyses, have been deposited in three data archives where they are available upon request: (*a*) CIPAC, Condor 214, Mexico 20, D. F.; (*b*) Children's Television Work-Shop, One Lincoln Plaza, New York, New York 10023; and (*c*) Hogg Foundation for Mental Health, University of Texas, Austin, Texas 78712.

TABLE 2

INTERCORRELATIONS AMONG PRETEST SCORES FOR 73 FOUR-YEAR-OLDS

Test	PN	PR	SP	PPT	HPC	PL	CS	PCO
General Knowledge	.67*	.42	.23	.12	.06	.31	.36	.21
Numbers (PN)		.27	.24	.08	.11	.38	.39	.21
Relations (PR)			−.06	.28	.17	.10	.11	.07
Embedded Figures (SP)				−.04	.12	.02	.16	.14
Parts of the Whole (PPT)					.00	.08	−.02	.10
Ability to Sort (HPC)						.17	.13	.13
Letters and Words (PL)							.31	.34
Classification Skills (CS)								.28
Oral Comprehension (PCO)								

* For 73 cases, correlations of .23, .30, and .38 are significantly different from zero at the .05, .01, and .001 levels, respectively.

category of tests is called independent-cognitive measures and consists of tests that have no direct or indirect relation to the stated goals of "Plaza Sesamo." Oral Comprehension is an example of such tests. One would expect to find the most significant differences between experimental and control groups on the content-achievement measures and the least significant differences on the independent-cognitive tests.

Intercorrelations among the nine major tests in these three different categories turned out to be of generally low order. With the exception of moderately high correlations for General Knowledge and Numbers (in the .60s), intercorrelations ranged from zero into the .30s and .40s. A sample matrix, the one for intercorrelations among pretest scores for all 4-year-olds, is presented in Table 2. The generally low level of intercorrelations indicates that differences between the experimental and control groups can be looked at independently for each of the measures without fear of undue repetition.

The second method of analysis consisted of a series of analyses of covariance in which the during-test measures and the posttest measures were adjusted statistically for regression upon the pretest measures. This analysis is especially powerful where significant correlations exist across time for a given measure. Of course, where there is no correlation between initial and later measures on a particular test, the analysis of covariance reduces to a simple analysis of variance of the posttest scores. Analysis of covariance adjusts for initial differences (even though these are essentially random), yielding estimates of the net gain made by the experimental children who viewed "Plaza Sesamo" as contrasted to the control cases who did not.

The third statistical method employed in analysis of the test data was analysis of variance in a repeated measures design to study trends in both experimental and control groups across time. Independent variables in the design were treatment (experimental vs. control), sex of child (male vs. female), and trials (pretest, during test, posttest)—a $2 \times 2 \times 3$ design. This analysis of variance design was employed separately for each of the three age groups. Still another analysis was undertaken in which the sex of the child was ignored and the three factors in the design consisted of age group (3, 4, and 5), treatment, and trial. And finally, a simple analysis of variance ignoring trials and dealing only with the posttest scores was undertaken for each of the three age groups separately, as well as the total sample combined, using only two independent factors—sex (male vs. female) and treatment (experimental vs. control).

The highlights of the findings from these several methods and analyses are presented for each of the three categories of tests, one test at a time.

Content-Achievement Tests

General Knowledge. The most rigorous test of gains made by the experimental groups who viewed "Plaza Sesamo," as compared to the control groups who did not, is

TABLE 3

POSTTREATMENT MEAN SCORES FOR EXPERIMENTAL AND CONTROL GROUPS ADJUSTED FOR REGRESSION ON INITIAL SCORES (GENERAL KNOWLEDGE)

Age group	Experimental (E)	Control (C)	E–C difference	Adjusted SD	Pre–post r
3-year-olds	14.7	12.0	2.7	5.4	.49
4-year-olds	26.2	18.9	7.3+	4.7	.63
5-year-olds	30.3	25.5	4.8+	4.9	.64
Combined	24.3	19.3	5.0+	5.5	.78

+ E–C difference significant at or beyond .001 level.

provided by the analysis of covariance. The results for General Knowledge are presented in Table 3. The 4-year-olds made the largest gain in General Knowledge, as evidenced by the superiority of the experimental over the control, a difference of 7.3 points. The 5-year-olds who viewed "Plaza Sesamo" also showed significant improvement, the experimental group receiving a mean score of 30.3, as compared to a mean of 25.5 for the control cases on the adjusted posttreatment scores. Only the 3-year-olds failed to show a significant difference between the experimental and control groups, although the small difference noted (2.7) is in the expected direction. Since the maximum score on General Knowledge is 37, the 5-year-olds who viewed "Plaza Sesamo" generally found the test very easy. The apparently greater improvement for the 4-year-olds, as contrasted to the 5-year-olds, may be partly due to a minor ceiling effect for the 5-year-old experimental children. In both cases, however, the amount of improvement for the experimental group is dramatic.

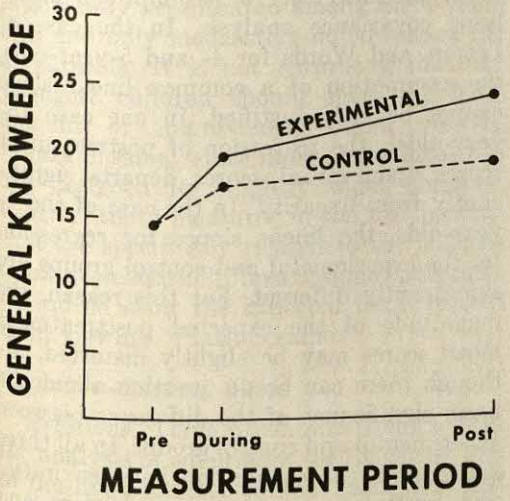


FIGURE 1. Treatment by trial interaction for General Knowledge.

It should also be noted in Table 3 that General Knowledge is fairly stable across time, yielding correlations ranging from .49 for 3-year-olds to .64 for 5-year-olds. The pre- and postcorrelation for the total sample of all ages combined is somewhat higher than for each of the ages separately (.78) because of the greater heterogeneity in the total sample since age is not controlled. The adjusted standard deviation of the posttreatment scores remains fairly constant across the three age groups, ranging from 4.7 for 4-year-olds to 5.4 for 3-year-olds.

Greater insight into the trends from the beginning to the end of the experiment can be gained by looking at the interaction of trial (the three points at which measurements were taken) with age group and treatment (experimental vs. control). Since the triple interaction involving all three factors is not significant, any trend differences between the experimental and control groups can be looked at for all three age groups combined. In a similar manner, different trends for each of the three age groups can be looked at without regard to whether the child is in the experimental or control group. The interaction of treatment by trial is presented in Figure 1. Here it can be seen that, although the experimental and control cases start out at essentially the same point, the experimental group increases more rapidly than the control group throughout the experiment.

The interaction of age by trial for experimental and control groups combined is presented in Figure 2. Gains made by the 3-year-olds are flatter than those for 4- and 5-year-olds. Both the 3- and 5-year-olds show slightly decelerating trends over time, while the 4-year-olds continue to show constant gains over time. While these absolute differences in slopes are not large, the interaction is highly significant.

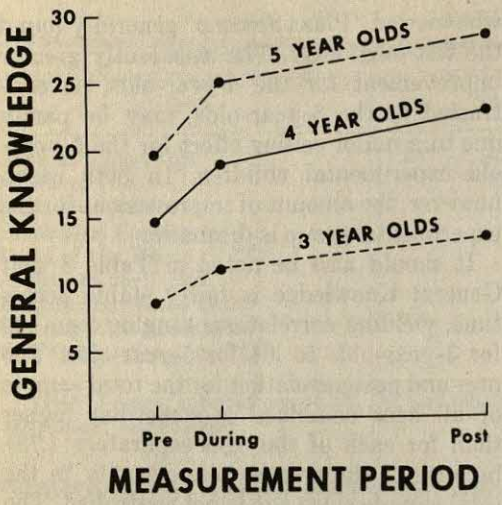


FIGURE 2. Age by trial interaction for General Knowledge.

The two-way analysis of variance for the combined age groups involving treatment and sex revealed no sex differences for General Knowledge, indicating that the major impacts of "Plaza Sesamo" upon performance on General Knowledge apply equally to both boys and girls.

Numbers. Results for the analysis of covariance for the Numbers Test are presented in Table 4. As in the case of General Knowledge, 4-year-olds show the greatest gain from viewing "Plaza Sesamo." Unlike General Knowledge, however, even the 3-year-olds show a significant gain in ability to cope with numbers. Results for the analyses of variance across trials involving Numbers proved to be essentially the same as the results already reported for General Knowledge. The experimental group showed a steady improvement at a much faster rate than did the control group; a significant age by trial interaction indicates that the 3-year-olds improved at a much slower rate than did the older children, and no sex differences were noted.

TABLE 4

POSTTREATMENT MEAN SCORES FOR EXPERIMENTAL AND CONTROL GROUPS ADJUSTED FOR REGRESSION ON INITIAL SCORES (NUMBERS)

Age group	Experimental (E)	Control (C)	E–C difference	Adjusted SD	Pre–post r
3-year-olds	17.0	12.6	4.4[+]	5.0	.32
4-year-olds	28.9	21.1	7.8[++]	6.9	.39
5-year-olds	35.5	29.3	6.2[++]	6.0	.56
Combined	27.9	21.5	6.4[++]	7.1	.71

[+] E–C difference significant at .01 level.
[++] E–C difference significant at or beyond .001 level.

TABLE 5

POSTTREATMENT MEAN SCORES FOR EXPERIMENTAL AND CONTROL GROUPS ADJUSTED FOR REGRESSION ON INITIAL SCORES (LETTERS AND WORDS)

Age group	Experimental (E)	Control (C)	E–C difference	Adjusted SD	Pre–post r
3-year-olds	10.4	7.5	2.9[+]	3.1	.39
4-year-olds	14.9	10.4	4.5[++]	3.9	.21
5-year-olds	17.8	12.7	5.1[++]	4.4	.16
Combined	14.7	10.3	4.4[++]	4.5	.43

[+] E–C difference significant at .01 level.
[++] E–C difference significant at or beyond .001 level.

Letters and Words. Table 5 contains results from the analysis of covariance for Letters and Words. This test proved to be much more difficult for many of the children than did either General Knowledge or Numbers. The adjusted posttreatment mean score for even the experimental 5-year-olds is only 17.8 out of a maximum score of 51. The computer program employed for analysis of covariance also tests for assumptions underlying covariance analysis. In the case of Letters and Words for 4- and 5-year-olds, the assumption of a common linear slope cannot be fully justified. In one case (5-year-olds), the regression of posttreatment scores upon initial scores departs significantly from linearity. In the case of the 4-year-olds, the linear slopes for regression for the experimental and control groups are significantly different. For this reason, the magnitude of the expected posttreatment mean scores may be slightly distorted, although there can be no question about the basic significance of the difference between experimental and control groups. In all three age groups, the correlation between initial and posttreatment scores on Letters and Words is so low (ranging from .16 to .39)

that very little adjustment has taken place in the posttreatment scores by use of regression techniques in analysis of covariance.

Greater insight into the differences between the experimental and control groups in their patterns of improvement over time can be gained by examining the interactions in the analysis of variance involving repeated measures. Lack of a significant triple-order interaction among age group, trials, and treatment indicates that the significant interaction of treatment by trials can be easily interpreted. The experimental group clearly improves across the three measurement periods while the control group does not. Both the experimental and control groups have means of approximately 8.6, initially, whereas the posttreatment means are 14.4 for experimental and 10.2 for control, a difference significant well beyond the .001 level. No significant differences were noted between boys and girls.

All three of the tests classified as content-achievement measures showed similar results in the analyses of covariance and variance. Although experimental and control groups started out at essentially the same place in the pretest measures, the experimental group quickly outdistanced the control in the first seven weeks of the experiment, maintaining and increasing their gains by the end of the six months. This widening gap between experimental and control cases is generally accentuated among the 4-year-olds and only moderately noted among the 3-year-olds. It is not surprising that the youngest children should show the least amount of improvement when viewing "Plaza Sesamo" since most of the measures are designed for slightly older children. Clearly, those measures in the test battery that are specifically related to the instructional segments of "Plaza Sesamo" programs do indeed show the expected improvement from viewing "Plaza Sesamo."

Cognitive-Content Tests

Relations. Table 6 contains results from the analysis of covariance for Relations, one of the five cognitive content measures used in the study. It is interesting to note that the correlation between initial and post-

TABLE 6

POSTTREATMENT MEAN SCORES FOR EXPERIMENTAL AND CONTROL GROUPS ADJUSTED FOR REGRESSION ON INITIAL SCORES (RELATIONS)

Age group	Experimental (E)	Control (C)	E–C difference	Adjusted SD	Pre-post *r*
3-year-olds	3.7	3.2	.5	1.4	.19
4-year-olds	5.6	4.0	1.6+	1.5	.01
5-year-olds	5.9	5.2	.7	1.7	.19
Combined	5.1	4.2	.9+	1.8	.27

+ E–C difference significant at .01 level.

treatment scores on Relations is essentially zero, ranging from .01 to .19 among the three age groups. Differences between the experimental and control groups were significant only for the 4-year-olds and then only at the .01 level.

Essentially the same results were obtained in the analysis of variance using repeated measures across the three testing periods; only the 4-year-olds show any significant advance in performance on Relations as a result of viewing "Plaza Sesamo." Probably the relatively insignificant results for Relations arise in large part from the unreliability of the test. The lack of any correlation across time suggests little test–retest stability. With only eight items in the test, it is not surprising that the gains are rather inadequately measured.

Parts of the Whole. The analysis of covariance for Parts of the Whole is summarized in Table 7. The differences between experimental and control groups are relatively small although in the expected direc-

TABLE 7

POSTTREATMENT MEAN SCORES FOR EXPERIMENTAL AND CONTROL GROUPS ADJUSTED FOR REGRESSION ON INITIAL SCORES (PARTS OF THE WHOLE)

Age group	Experimental (E)	Control (C)	E–C difference	Adjusted SD	Pre-post *r*
3-year-olds	3.4	2.9	.5	1.5	.21
4-year-olds	5.0	4.4	.6	2.1	−.02
5-year-olds	6.7	5.1	1.6+	2.5	.06
Combined	5.2	4.2	1.0++	2.3	.14

+ E–C difference significant at .05 level.
++ E–C difference significant at .01 level.

TABLE 8

POSTTREATMENT MEAN SCORES FOR EXPERIMENTAL AND CONTROL GROUPS ADJUSTED FOR REGRESSION ON INITIAL SCORES (ABILITY TO SORT)

Age group	Experimental (E)	Control (C)	E–C difference	Adjusted SD	Pre-post r
3-year-olds	3.4	2.9	.5	1.7	.21
4-year-olds	7.7	4.7	3.0+	2.6	.07
5-year-olds	9.6	6.4	3.2+	3.2	.34
Combined	7.5	4.9	2.6+	3.1	.32

+ E–C difference significant at or beyond .001 level.

tion. Only the 5-year-olds, when considered alone, show a difference that reaches statistical significance. When all three groups are combined, however, the greater stability of results with a much larger number of cases yields a reliable difference in favor of the experimental groups. It should be noted, however, that the difference in absolute terms is very small, amounting to only 1 item on the average. Since there are only 10 items in this test, these results suggest that the stability of measurement and the clarity of results would be greatly enhanced by increasing the length of the test and obtaining more discriminating items for any future revision.

In the complex analysis of variance using repeated measures, Parts of the Whole shows significant interactions between both age groups and treatment, on the one hand, and trials on the other. As would be expected from the covariance analysis, the experimental group shows a sharper rise in mean score across time than does the control group. The increase is appreciably greater for 5-year-olds than for either of the younger groups. No sex differences were discovered.

Ability to Sort. As indicated in Table 8, after viewing "Plaza Sesamo" both 4-year-olds and 5-year-olds show significant gains in their ability to sort different objects, things, or persons. Only the 3-year-olds failed to show any significant difference, although a slight trend appears in the expected direction. When all three groups are combined, the results are highly significant, of course. Again, it should be noted that there is a relatively low correlation between pre- and posttest measures, in this case especially for 4-year-olds. In the case of the 4- and 5-year-olds, a net gain of 3 points for the experimental over the control groups is indeed an appreciable one.

Greater insight into the meaning of these results can be obtained by closer examination of the analysis of variance using repeated measures. Highly significant interactions were obtained for age by trial and treatment by trial, although the triple interaction was not significant. Consequently, generalizations can be made for all three age groups with respect to the superior performance of the experimental groups. The combined mean for all three age groups in the pretest period is about 3.8 for both experimental and control cases. The gap between experimental and control groups widens seven weeks later to 5.5 and 4.8 for experimental and control groups, respectively. By the end of the experiment, the children who viewed "Plazo Sesamo" received a mean of 7.3 as contrasted to only 4.9 for the control cases. It is interesting to note that the experimental group steadily improved throughout the experiment while the control group showed only minor improvement (probably due to test–retest adaptation) in the early stages of the experiment, showing no improvement thereafter. Again, no differences were noted between boys and girls regardless of age group.

Classification Skills. Results of the analysis of covariance are summarized for Classification Skills in Table 9. All three age groups show significant differences between the experimental and control groups. The greatest difference appears among the 4-year-olds, a finding consistent with results reported for other measures. The interactions of treatment with age and trial in the analysis of variance using repeated measures were also highly significant. The experimental group shows a steady and marked increase across the testing periods while the increase for the control group is rather slight. No differences were noted between boys and girls.

Embedded Figures. The last of the five

TABLE 9

POSTTREATMENT MEAN SCORES FOR EXPERIMENTAL AND CONTROL GROUPS ADJUSTED FOR REGRESSION ON INITIAL SCORES (CLASSIFICATION SKILLS)

Age group	Experimental (E)	Control (C)	E–C difference	Adjusted SD	Pre-post r
3-year-olds	6.9	4.9	2.0++	2.3	−.01
4-year-olds	11.9	8.7	3.2+++	2.9	.37
5-year-olds	14.4	12.1	2.3+	3.6	.36
Combined	11.5	8.7	2.8+++	4.0	.45

+ E–C difference significant at .05 level.
++ E–C difference significant at .01 level.
+++ E–C difference significant at or beyond .001 level.

measures classified as cognitive-content measures is Embedded Figures, the preschool adaptation of Witkin's Embedded Figures Test. It is less closely related to the content of "Plaza Sesamo" than any of the other tests in this cognitive-content group, although some of the instructional sequences in "Plaza Sesamo" appear closely related to the type of ability measured by Embedded Figures. Table 10 summarizes results for the analysis of covariance.

Unlike most of the other tests, Embedded Figures shows a relatively high degree of stability across the six-month period from pre- to posttreatment measurement—correlations ranging from .45 to .64. Consequently, the covariance analysis is appreciably more precise than would be a simple analysis of variance of the posttreatment scores alone. Only the 4-year-olds show a significant difference between the experimental and control groups and even this difference is significant at only the .05 level. Among the 5-year-olds, there appears to be no difference between the adjusted mean values for the experimental and control groups. For the 3-year-olds, the difference is somewhat greater (3.7 points) and almost reaches significance ($p = .06$).

Since there are 24 items in the test, it is obvious that Embedded Figures is much too easy for the 5-year-olds, probably accounting for the lack of difference between the experimental and control groups. The large standard deviation (6.8) for adjusted posttreatment scores among the 3-year-olds suggests that the individual differences among young children on this test are so great that even a difference of 3.7 points, as obtained between experimental and control groups, is not sufficient to be impressive. When the analyses of covariance for posttest scores regressed upon during-treatment scores and for during-test scores regressed upon initial scores are examined, it is clear that among the 3-year-olds what little gain there is takes place rather late in the experiment. No significant sex differences were noted. The above results suggest that Embedded Figures would be a promising instrument for future studies of this type provided it is revised and extended to overcome the deficiencies noted above.

While results for the cognitive-content tests are not quite as striking as those for the more specific content-achievement tests, in every one of the five tests at least one age group showed significantly greater improvement over time for those children who viewed "Plaza Sesamo" than for those who did not. The measures most consistent in this regard were the two tests dealing with different aspects of categorizing behavior—Ability to Sort and Classification Skills. The results might have been more striking for Parts of the Whole and for Relations if these two tests were not so unstable across time. Since both of them can be improved considerably by lengthening the tests and including more discriminating items, the cognitive abilities represented by them should be studied carefully in future evaluations of "Plaza Sesamo." Even Embedded

TABLE 10

POSTTREATMENT MEAN SCORES FOR EXPERIMENTAL AND CONTROL GROUPS ADJUSTED FOR REGRESSION ON INITIAL SCORES (EMBEDDED FIGURES)

Age group	Experimental (E)	Control (C)	E–C difference	Adjusted SD	Pre-post r
3-year-olds	13.0	9.3	3.7	6.8	.45
4-year-olds	19.4	17.4	2.0+	3.7	.48
5-year-olds	20.9	20.4	.5	3.0	.64
Combined	18.0	16.1	1.9+	5.0	.68

+ E–C difference significant at .05 level.

TABLE 11

POSTTREATMENT MEAN SCORES FOR EXPERIMENTAL AND CONTROL GROUPS ADJUSTED FOR REGRESSION ON INITIAL SCORES (ORAL COMPREHENSION)

Age group	Experimental (E)	Control (C)	E-C difference	Adjusted SD	Pre-post r
3-year-olds	13.6	10.5	3.1+	5.4	.44
4-year-olds	23.4	18.1	5.3++	6.4	.22
5-year-olds	26.2	23.8	2.4+	4.2	.59
Combined	21.6	17.9	3.7+++	6.1	.67

+ E–C difference significant at .05 level.
++ E–C difference significant at .01 level.
+++ E–C difference significant at or beyond .001 level.

Figures, the test in this category most remotely associated with "Plaza Sesamo" content and goals, yielded significant differences for the 4-year-olds (and nearly significant differences for the 3-year-olds).

Independent-Cognitive Measures

Oral Comprehension. The one test among the many used for evaluation of "Plaza Sesamo" that, on the face of it, was completely unrelated to the content and goals of the television program was Oral Comprehension, the test developed for use with young bilingual children. As can be seen in Table 11, the results for all three age groups in the analysis of covariance were significant. The 4-year-olds showed the greatest gains from viewing "Plaza Sesamo," a net difference of 5.3 points advantage over the controls. The test–retest stability ranges from .22 to .59 for the three age groups, indicating that the analysis of covariance adds significantly to the precision of the analysis. Results from the analysis of variance using repeated measures are essentially the same as those obtained for most of the other measures in this study.

It is interesting to note that highly significant gains were made by the experimental group, as contrasted to the control group, on Oral Comprehension even though it has no relation to the curriculum and goals of "Plaza Sesamo." This important finding suggests that there may well be general gains of a cognitive nature as a result of viewing "Plaza Sesamo." Efforts should be made to explore other cognitive, perceptual, and linguistic measures that could be incorporated into future evaluations of "Plaza Sesamo."

Other Measures

In addition to the tests given individually to both experimental and control children, observational ratings on attention were obtained on a random sampling basis for every child in the experimental group in two day-care centers. The number of absences was also recorded for every child in the study. Because of the fact that attention ratings were only available for the experimental group, no analyses of variance or covariance could be undertaken. The highly skewed nature of the distribution of absences also precluded such analyses.

Within the combined experimental group, the degree of attention to "Plaza Sesamo" correlated significantly with pretest Numbers (.32), during-test Classification Skills (.32), and six tests from the posttreatment battery—General Knowledge (.32), Numbers (.35), Parts of the Whole (.30), Ability to Sort (.49), Classification Skills (.33), and Oral Comprehension (.41). The larger number of appreciable correlations in the posttreatment test battery is to be expected since attention (or lack of it) would have a cumulative effect over the six months of the "Plaza Sesamo" programs. Clearly, degree of attention to what is going on is important to measure.

The number of times a child was absent also influenced his performance on some of the tests in the experimental group, frequently yielding correlations in the .20s with posttest measures of Oral Comprehension, Classification Skills, Letters and Words, Relations, and Numbers. Children with a large number of absences did less well on the posttreatment test battery than did children who attended regularly. These results are not surprising and indicate, as in the case of attention, that the number of absences should be carefully recorded in any evaluation study of this type. The correlation between degree of attention and number of absences was −.26 (significant at the .05 level).

The outcome of this experimental study of preschool children, half of whom viewed "Plaza Sesamo" and half of whom did not, clearly indicates that significant gains are made in a number of cognitive and perceptual areas by those children who faithfully watch the television program. While such a finding is not surprising for those areas of specific achievement related to the curriculum content of the television programs, it is indeed significant that even measures of cognitive ability seemingly unrelated to the main thrust of the curriculum show measurable improvement in children who viewed the program as contrasted to those who did not.

REFERENCES

Ball, S., & Bogatz, G. A. *The first year of Sesame Street: An evaluation.* Princeton, N. J.: Educational Testing Service, 1970.

Bogatz, G. A., & Ball, S. *The second year of Sesame Street: A continuing evaluation.* Vols. 1 & 2. Princeton, N. J.: Educational Testing Service, 1971.

Coates, S. *Preschool Embedded Figures Test.* Palo Alto, Calif.: Consulting Psychologist's Press, 1971.

Cook, T. *Sesame Street, a history of an evaluation. A report to the Russell Sage Foundation,* 1972. (Mimeo)

Manuel, H. T. *Comprension del Lenguaje Oral Edicion Revisada, Nivel 1* Austin, Texas: Guidance Testing Associates, 1972.

(Received December 26, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 644-656

ACQUISITION PROCESSES AND RESILIENCE UNDER VARYING TESTING CONDITIONS FOR STRUCTURALLY DIFFERENT PROBLEM-SOLVING PROCEDURES[1]

RICHARD E. MAYER[2]

Indiana University

Binomial probability problems were taught by emphasizing either calculating with a formula (Sequence F) or meanings of the variables in the formula (Sequence G). Two main experiments involved a multileveled transfer posttest administered after the subject read some of his instructional booklet and variational testing conditions (e.g., open versus closed book, speed versus power). A Treatment × Posttest interaction (TPI) resulted in which Sequence G subjects excelled on interpretive items while Sequence F subjects excelled on near transfer and the Treatment × Posttest interaction did not differ reliably among different points in learning or under different testing conditions. Results indicated the importance of a subject's initial assimilative set in acquiring new knowledge and the apparent resilience of the acquired structure.

Previous studies (Egan & Greeno, 1973; Mayer & Greeno, 1972) have suggested that teaching subjects to solve mathematical problems by different instructional methods may result in learning outcomes which differ in structural or qualitative ways. This inference was indicated by a pattern of posttest performance in which subjects in one instructional group excelled on one kind of transfer posttest item and subjects in another group excelled on another kind, producing a disordinal interaction called Treatment × Posttest interaction or TPI.

Development of structural difference. A major new question dealt with in the present study is, How do the cognitive structures that support these different patterns of transfer performance develop over the course of learning? Or, put another way, How can we characterize the acquisition processes for structurally different learning outcomes? At least two kinds of theories of the acquisition process have been proposed: (*a*) A fairly straightforward idea, one that follows from asking "how much" is learned, is that apparent differences in what is learned are due to some subjects acquiring more of one kind of content and less of another relative to other subjects. (*b*) A more complex proposition, one that follows from asking "what is learned" (e.g., Roughead & Scandura, 1968), is that different kinds of learning outcomes are due to acquisition processes in which the same content material is encoded within different assimilative sets by different subjects. Although the first proposal requires only an analysis of the amount of material transmitted, most recent educational theories have relied on modified versions of the second proposal in which the subject's cognitive activity or receptive set during learning as

[1] This research was supported by the Advanced Research Projects Agency under Contract F44620-73-C-0019 with the Human Performance Center, University of Michigan. This report is based on a dissertation submitted to the University of Michigan in partial fulfillment of requirements for the degree of Doctor of Philosophy in Psychology. I appreciate support from the members of the dissertation committee: James G. Greeno, chairman, and Joseph Payne, Robert Bjork, and Steven Kaplan. The assistance of Wilbert J. McKeachie is also appreciated.

[2] Requests for reprints should be sent to Richard E. Mayer, who is now with the Department of Psychology, Indiana University, Psychology Building, Bloomington, Indiana, 47401.

well as the presented material must be analyzed.

The present study attempts to provide information concerning the nature of the acquisition process for mathematical knowledge and to provide more direct information on the hypothesis stated by Mayer and Greeno (1972):

> different instructional procedures could activate different aspects of existing cognitive structure. And since the outcome of learning is jointly determined by new material and the structure to which it is assimilated, the use of different procedures could lead to the development of markedly different structures during the learning of the same concept [p. 165].

Resilience of structural differences. A second new question dealt with in the present study is, Once learning outcomes have been established, can the pattern of transfer performance be altered by testing manipulations such as open book versus closed book, or speed versus power conditions? In other words, how resilient are structurally different learning outcomes to varying testing conditions? Such manipulations represent an attempt to force subjects who have learned by different instructional methods to processes "what is learned" in the same way. If structural differences "disappear" under these circumstances, that is, the Treatment × Posttest interaction is not present, the importance of the original teaching method in establishing the structure of learning outcomes would be diminished.

Present study. In the present experiments, the concept of binomial probability was taught using expository, four-lesson teaching booklets and two teaching methods that differed in sequencing and emphasis. One instructional method (Sequence F) began each lesson with a formal statement of the rule or subrule and explained component variables only within the context of calculating with the formula. The other method (Sequence G) began each lesson by attempting to relate component variables of the formula to the subject's general experience, for example, trials, outcomes, and successes, before presenting any formal statement of the rule. Thus, Sequence F attempted to activate a narrow range of subject's experience such as performing arithmetic computations, while Sequence G attempted to activate a broader receptive set in the subject including a wide range of general experience with probabilistic situations. These procedures represent an attempt to provide subjects with what Ausubel (1968) has termed "rote" versus "meaningful learning sets" or what Greeno (1972) has called "algorithmic" versus "propositional knowledge." Learning was assessed by a multileveled transfer posttest which contained both near (i.e., items very similar to those in the training booklets) and far (i.e., items requiring interpretation) items.

To provide information on the acquisition question, the posttest was administered at three points in learning for subjects in both instructional groups. Amount 1 subjects were tested after the first two lessons (introduction and combinations), Amount 2 subjects after three lessons (introduction, combinations, and joint probability), and Amount 3 subjects after all four lessons (introduction, combinations, joint probability, and binomial probability) had been presented. Both Experiment 1 and Experiment 2 used this procedure, although the ordering of booklet lessons differed. In addition, a supplemental study was conducted in which the subject, instead of taking a transfer posttest, was asked to reproduce what he had just read as if he were explaining it to a naive learner.

To provide information on the resilience question, the conditions of testing were varied. In Experiment 1, the testing variable was memory support with some subjects tested under open-book and others under closed-book conditions. In Experiment 2, the testing variable was time stress with some subjects tested under speed (time limit) and some under power (no-limit) conditions. In addition, a supplemental study was conducted which involved a two-day retention interval.

Experiment 1

This study attempted to assess both the consequences of varying the sequencing and amount of instruction for a mathematical concept and the presence of memory

support during testing. In each instructional sequence, subjects were given varying amounts of instruction. All subjects took the same 30-item posttest, containing several kinds of test items, and the condition of testing was varied.

Method

Subjects and Design

Subjects were 117 University of Michigan students who had volunteered to participate in psychological experiments at the Human Performance Center for pay. Nine subjects served in each cell of a 2 × 3 × 2 factorial design, with a thirteenth group of subjects serving as controls. The factors were method of instruction (Group F or Group G), amount of instruction (Amount 1, 2, or 3), and testing condition (open book or closed book).

Materials

The two instructional texts (Sequence F and Sequence G) were incorporated into 2 four-lesson typewritten booklets with one to three pages in each lesson. The booklets were similar to those used in earlier studies (Mayer & Greeno, 1972) and have been reproduced in full elsewhere (Mayer, 1973).

Instruction in Sequence F (for formula) consisted of the following four lessons: (*a*) an introduction, formally presenting the binomial formula and showing how it could be broken down into three smaller algorithms; (*b*) combinations, emphasizing how to obtain a value for $N!/[(N - R)!R!]$ (where N is number of trials and R is number of successes); (*c*) joint probability, emphasizing how to obtain a value for $P^R(1 - P)^{N-R}$ (where P is probability of successes); and (*d*) binomial probability, emphasizing how to find the product of the above values to obtain the final answer. Thus, Sequence F began each lesson with the formal presentation of the formula or subformula, and component concepts (e.g., trial, failure, failure probability) were explained only in relation to calculating with the formula. Throughout instruction in Sequence F, the amount of interpretation of variables was minimized, and emphasis was on numerical calculation.

The instructional booklet for Sequence G (for general concepts) consisted of the following four lessons: (*a*) an introduction, presenting definitions of variables (e.g., number of trials, number of successes, etc.) in relation to general experience; (*b*) combinations, reemphasizing the relevant concepts before presenting the subformula for combinations; (*c*) joint probability, reemphasizing relevant concepts before presenting the subformula for joint probability; and (*d*) a binomial probability, which tied the subformulas together conceptually before presenting the full binomial formula. Thus, Sequence G began each lesson with individual variables and developed from parts to the whole concept. Through the instruction in Sequence G, emphasis was on explanations of the meanings of concepts, and calculation was discussed only when full conceptual explanation had been given.

A test set consisted of 30 typewritten cards representing five problem types, two problem formats, and three problem content areas. The 5 × 2 × 3 design yielded 30 cells, each represented in one of the 30 test items.

The first transfer dimension was the type of test item. Familiar problems (F type) were presented in the same way as example problems during training; transformed problems (T type) required a transformation, usually of an algebraic nature, to be put into familiar form; so-called Luchins problems (L type) presented a complicated-looking situation which could be solved quite easily if the subject would take a moment to "think"; question items (Q type) asked a question about the variables in the formula rather than requiring a solution value; and unanswerable problems (U type), although looking very much like F-type problems, presented either insufficient or inconsistent information.

The second dimension was test problem format. Formula problems (F format) were stated in terms of N, R, and P—the formal notation used in presenting the formulae in the teaching booklets; story problems (S format) were stated in terms of some situation not discussed in the teaching booklets such as sampling peanuts from a barrel in which some portion is rotten.

The third dimension was the content of test problems. Combinations problems (C context) asked for or dealt with C (N, R) or the number of combinations; joint probability problems (J content) asked for or dealt with $P^R \times (1 - P)^{N-R}$ or the probability of a specific sequence; binomial probability (B content) asked for or dealt with the theme of the teaching booklets, finding P (R, N) or the probability of R successes in N trials.

Examples of the test problems and answers are published elsewhere (Mayer, 1973).

Additional materials were a subject record consisting of questions to determine the extent of the subject's experience in statistics and probability and a pretest designed to determine whether the subject had sufficient computational skill to master the material in the teaching booklets.

Procedure

Subjects were run in groups averaging four per session. First, the subject record and pretest were administered. Subjects indicating no relevant experience with the binomial but making no more than three computational errors were randomly assigned to treatment groups.

The experimenter read the instructions and again asked the subject to indicate familiarity with the binomial. Then, each subject was given the appropriate teaching booklet. Subjects were instructed to read their booklets silently and at

their own rate and to try to understand the explanations and examples. Subjects were allowed to take notes or figure on a blank piece of paper. Subjects were told they would have 20 minutes. All finished within that time limit. Subjects who finished earlier were asked to sit quietly until the test began.

Immediately following the reading period, the experimenter collected the teaching booklets and notes (except for memory support condition subjects), read the instructions for the test, and wrote the formula for P (R, N) on the chalkboard. Each subject was given a pile of 30 problem cards face down and a blank answer sheet. On the experimenter's signal, the subject was to turn up the first card, copy its code number, show his work and circle his final answer, and then go on to the next card. There was no time limit, but all subjects were told that once they began a new card they could not go back to work on any previous card. All subjects were told to write "no answer" if they felt a problem was unanswerable.

Nine different orders of presenting the test cards were constructed. The orders were random except for the constraint that, of the nine orderings, each item had to appear in the first one third (i.e., 1–10) in three of the sets, in the second one third (i.e., 11–20) in another three of the sets, and in the last one third (i.e., 21–30) in the other three sets.

Results

The posttest performance of subjects who passed the pretest but expressed no familiarity with the binomial was scored with each item marked either correct or incorrect. Answers were marked correct if they were in proper form even though computation may have been incorrect or not carried out. The control subjects performed at very low levels, that is, an average of less than 10% correct, thus indicating that the experimental treatments had a substantial effect. An analysis of variance was performed on the data of all experimental subjects. There was only a marginal difference in the overall performance of the two experimental groups ($F = 2.92$, $df = 1/96$, $p < .10$), thus frustrating the question of "which method is best."

Development of Structural Differences

Figure 1 shows the proportion correct response for the two instructional groups by posttest item (i.e., Treatment × Posttest interaction) for each of three amounts of instruction and for each of three posttest dimensions. For each of the three partitions

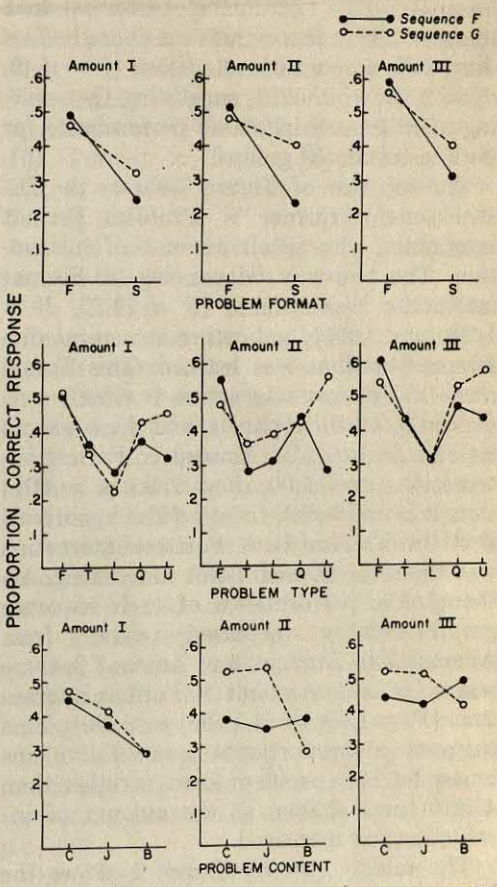


FIGURE 1. Treatment × Posttest interaction by amount of instruction for three posttest dimensions in Experiment 1. (Problem format: F = formula, S = story; problem type: F = familiar, T = transformed, L = Luchins, Q = question, U = unanswerable; problem content: C = combinations, J = joint probability, B = binomial probability.)

of the posttest set (i.e., by format, by type, or by content), there is reason to suspect that the final outcome (i.e., shown in the right panels) of learning differed between subjects in Sequence F and Sequence G, in a manner established in previous studies (Mayer & Greeno, 1972); that is, Sequence F subjects excelled on near transfer and Sequence G on far transfer.

In order to determine how these final differences in cognitive structure developed over the course of learning, it is useful to compare the pattern of Treatment × Posttest interaction in the left panel (after two lessons) to the center panel (after three

lessons) to the right panel (after all four lessons). There is a significant overall effect due to the amount of instruction ($F = 4.49$, $df = 2/29$, $p < .025$), suggesting that adding more lessons increased performance for both instructional groups.

The top row of Figure 1 shows the Instructional Sequence × Problem Format interaction after each amount of instruction. The two-way Sequencing × Format interaction was reliable ($F = 14.72$, $df = 1/96$, $p < .001$), indicating structural differences in what was learned. This Treatment × Posttest interaction is clearly evidenced in all three panels, and there was no reliable Sequence × Amount × Format interaction ($F = 1.00$, $df = 2/96$, $p > .20$); thus it is impossible to reject the hypothesis that the Treatment × Posttest interaction was the same at each point in learning. Although the performance of each sequence group reliably increased overall from Amount 1 to Amount 2 to Amount 3, there was no reliable Amount × Format interaction ($F = 1.00$, $df = 2/96$, $p > .20$). This suggests a proportional quantitative increase for both problem formats rather than a structural change as the amount of instruction was increased.

The middle row of Figure 1 shows the Instructional Sequence × Problem Type interaction after each amount of instruction. The two-way Sequencing × Type interaction was reliable ($F = 6.28$, $df = 4/384$, $p < .001$), suggesting structural differences similar to those found in earlier studies. Even after Amount 1, there is evidence of this characteristic Treatment × Posttest interaction with Sequence F slightly ahead for F type and Sequence G superior on Q type and U type. This pattern seems fairly consistent across all three amounts of instruction, and there was no reliable Sequence × Type × Amount interaction ($F = 1.03$, $df = 8/384$, $p > .20$); thus it is impossible to reject this hypothesis. Although the performance of each treatment group reliably increased overall from Amount 1 to Amount 2 to Amount 3, there was no reliable Amount × Type interaction ($F = 1.05$, $df = 8/384$, $p > .20$), again suggesting that the increase was uniform across all problem types. These findings are consistent with the notion that the structure of learning outcomes remained constant across all three amounts of instruction.

The bottom row of Figure 1 shows the Sequence × Problem Content interaction after each amount of instruction. The two-way Sequence × Content interaction was significant ($F = 10.19$, $df = 2/192$, $p < .001$), again indicating a difference in the structure of what was learned. There was no reliable Treatment × Content × Amount interaction ($F = 1.77$, $df = 4/192$, $p > .15$); thus, it is not possible to reject the notion that the structural differences among the instructional groups remained constant at each of the three points in learning which were tested. A marginally reliable Amount × Problem Content interaction ($F = 2.54$, $df = 4/192$, $p \approx .050$) suggests that, not only did performance increase overall with the addition of more sections, but it also increased disproportionately more, as might be expected, on the content of the material covered.

In summary, the three possible Sequence × Posttest × Amount interactions represent separate tests of the "acquisition question." In all three cases, there was a reliable two-way Treatment × Posttest interaction, but there was no evidence of any difference in the Treatment × Posttest interaction among Amount 1, Amount 2, and Amount 3. These results suggest that structural differences occurred early in learning and remained fairly constant throughout.

Resilience of Structural Differences

This study was also intended to provide some information about the resilience of the Treatment × Posttest interaction under varying testing conditions. Figure 2 shows the proportion correct response for the two instructional sequences by posttest item (i.e., Treatment × Posttest interaction) for each of two levels of memory support and for each of three posttest dimensions. The performance for the instructional sequence groups under memory-support or open-book (left panel) and no memory-support or closed-book (right panel) conditions is given by problem format in the top row, by problem type in the middle row, and by problem content in the bottom row. Open-

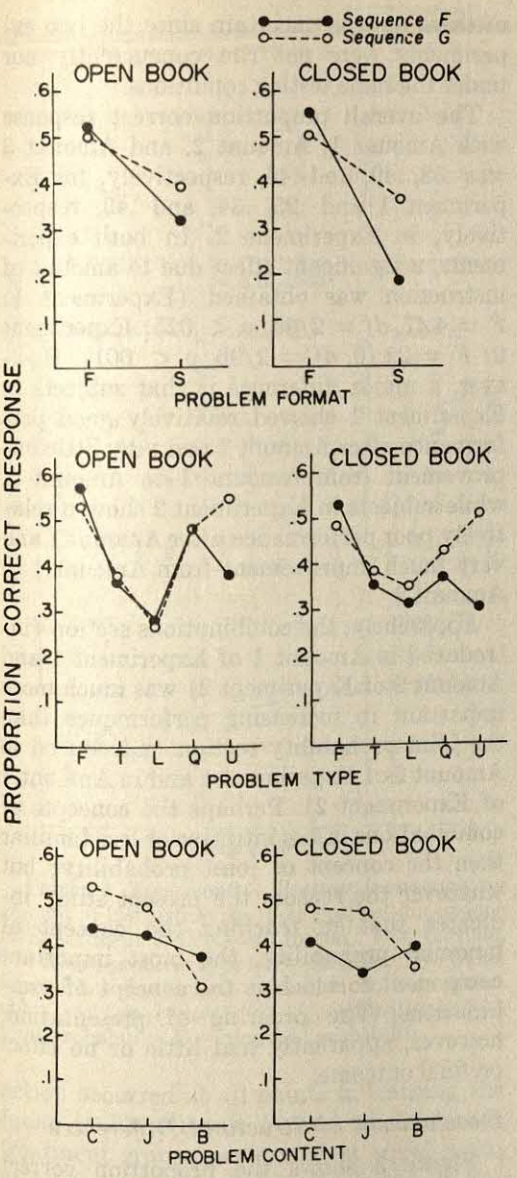


FIGURE 2. Treatment × Posttest interaction by testing condition for three posttest dimensions in Experiment 1. (Problem format: F = formula, S = story; problem type: F = familiar, T = transformed, L = Luchins, Q = question, U = unanswerable; problem content: C = combinations, J = joint probability, B = binomial probability.)

book testing seems to have helped to slightly increase posttest scores overall for both instructional groups, although the increase failed to reach a statistically significant level ($F = 1.44, df = 1/96, p > .20$).

For both instructional groups, there was some hint that memory support helped to disproportionately increase performance on harder problems—far transfer items—especially for S-format relative to F-format items. Marginally significant Type × Testing Condition ($F = 2.12$, $df = 4/334$, $p < .10$), Format × Testing Condition ($F = 4.79$, $df = 1/96$, $p < .05$), and Content × Testing Condition ($F = 2.57$, $df = 1/192$, $p < .10$) interactions offer only weak support for this observation.

An analysis of resilience of structural differences under varying testing conditions is made uncertain by the fact that there was no evidence that the memory-support manipulation produced a reliable effect. However, a comparison of right and left panels in each row of Figure 2 indicates that the Treatment × Posttest interaction was present under both open-book and closed-book conditions, and this observation is consistent with a failure to obtain a reliable three-way interaction among instructional sequence, type of posttest item, and testing condition ($F < 1.00$, $df = 4/184$, $p > .20$), among sequence, problem type, and testing condition ($F = 2.79$, $df = 1/96$, $p > .10$), or among sequence, problem content, and testing conditions ($F < 1.00$, $df = 2/192$, $p > .20$).

Open-book testing seems to have very little effect on either the amount or the structure of problem-solving performance. The fact that the test followed learning almost immediately and that the binomial formula was on the chalkboard for all subjects may have reduced the impact of open-book testing. Apparently, subjects in the closed-book condition were able to generate problem solutions quite well with only their existing knowledge and the formula on the chalkboard. These results suggest that the structural differences obtained do not depend on subjects having to remember material. Apparently failures in the test are due to failures to understand things rather than failures to retain information.

EXPERIMENT 2

Experiment 2 was intended to replicate and extend the results of Experiment 1. The main changes in Experiment 2 were that the ordering of instructional lessons was introduction, joint probability, combinations,

and binomial probability, and the testing condition variable was speed or time stress (limit of 90 seconds per problem) versus power or no time stress (no limit). In Experiment 2, no subjects received memory support during testing.

Method

As in Experiment 1, 9 subjects served in each cell of a 2 × 3 × 2 design with the factors being instructional sequence (Sequence F and Sequence G), amount of instruction (Amounts 1, 2, and 3), and testing condition (time stress and no time stress). The 108 subjects were recruited from a pool of University of Michigan students who had volunteered to participate in psychological experiments at the Human Performance Center for pay.

Because of the change in ordering, subjects receiving Amount 1 studied the introduction section and joint probability section; subjects given Amount 2 had the introduction followed by joint probability and combinations, and subjects given Amount 3 received all four sections in the order: introduction, joint probability, combinations, and binomial probability.

The transfer test performance was scored and analyzed as in Experiment 1. As in Experiment 1, there was no reliable difference in overall performance between the two treatment groups ($F = 1.80$, $df = 1/96$, $p > .15$).

Results and Discussion

Effect of Ordering

A variable that was changed from Experiment 1 to Experiment 2 was the ordering of the parts or sections in the instructional booklets. For example, Amount 1 contained information about an introduction and combinations in the ABC ordering of Experiment 1, but Amount 1 contained information about an introduction and joint probability in the BAC ordering of Experiment 2. Some indication of the effect of ordering and the respective importance of the combinations and joint probability sections is provided by comparing performance across experiments, although interpretation of these data is uncertain since the two experiments were not run concurrently nor under the same testing conditions.

The overall proportion correct response with Amount 1, Amount 2, and Amount 3 was .38, .40, and .46, respectively, for Experiment 1 and .25, .34, and .43, respectively, in Experiment 2. In both experiments, a significant effect due to amount of instruction was obtained (Experiment 1: $F = 4.47$, $df = 2/96$, $p < .025$; Experiment 2: $F = 22.10$, $df = 2/96$, $p < .001$). However, a major difference is that subjects in Experiment 1 showed relatively good performance after Amount 1 and very little improvement from Amount 1 to Amount 2, while subjects in Experiment 2 showed relatively poor performance after Amount 1 and very much improvement from Amount 1 to Amount 2.

Apparently, the combinations section (introduced in Amount 1 of Experiment 1 and Amount 2 of Experiment 2) was much more important in increasing performance than the joint probability section (introduced in Amount 2 of Experiment 1 and in Amount 1 of Experiment 2). Perhaps the concepts of combinations is less intuitive or less familiar then the concept of joint probability; but whatever the reason, the present study indicates that in teaching the concept of binomial probability, the most important component to teach is the concept of combinations. The ordering of presentation, however, apparently had little or no effect on final outcome.

Development of Structural Differences

Figure 3 shows the proportion correct response for the two instructional groups by posttest item (i.e., Treatment × Posttest interaction) for each of three amounts of instruction and for each of three posttest dimensions. As in Experiment 1, and as can be seen in the right panels of the three rows in Figure 3, the characteristic patterns of the Treatment × Posttest interaction were present, suggesting that the final learning outcomes of subjects in Sequence F and Sequence G were structurally different.

Again, special attention should be paid to whether the Treatment × Posttest inter-

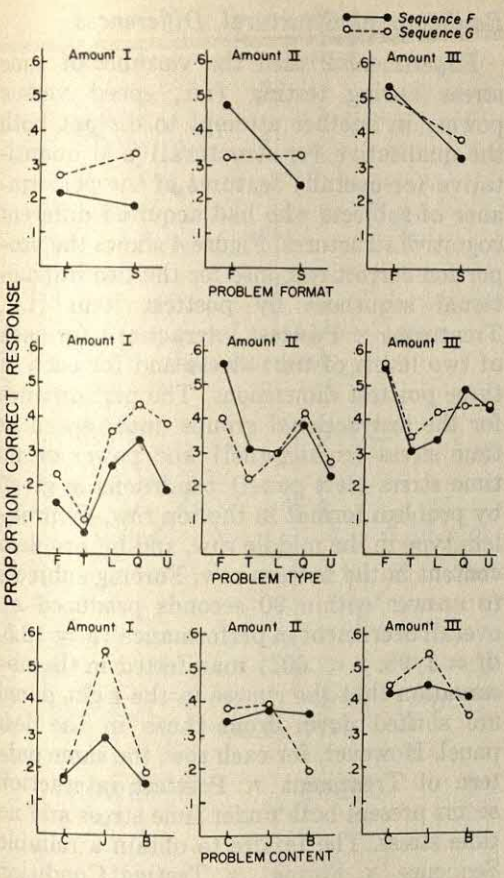


FIGURE 3. Treatment × Posttest interaction by amount of instruction for three posttest dimensions in Experiment 2. (Problem format: F = formula, S = story; problem type: F = familiar, T = transformed, L = Luchins, Q = question, U = unanswerable; problem content: C = combinations, J = joint probability, B = binomial probability.)

action occurred at all points in learning (as found in Experiment 1) and whether both treatment groups performed at equal levels at each point in learning (as found in Experiment 1).

The top row of Figure 3 shows the performance of the two instructional groups by problem format at each of three points in learning. The two-way Sequence × Format interaction—in which Sequence F subjects excelled on F-format and Sequence G subjects performed better on S-format items—was reliable ($F = 21.65$, $df = 1/96$, $p < .001$), again suggesting a structural difference in what was learned by subjects in the two treatment groups. The Sequence × Amount interaction ($F = 3.30$, $df = 2/96$, $p < .05$) was manifested in the fact that Sequence G outperformed Sequence F for both formats after Amount 1 but evened out to the usual Treatment × Posttest interaction after Amounts 2 and 3. Unlike Experiment 1, apparently, lack of specific content material was more detrimental for Sequence F subjects relative to Sequence G.

In addition, the same pattern of Treatment × Posttest interaction was obtained at each of three points in learning, although the interaction was not disordinal after Amount 1 largely due to the poor performance of subjects in Sequence F. The general thrust of these results seems to be as follows: with Sequence F, there is so little learned in Amount 1 that subjects can be said to have little or no structure, and when structure begins to develop, it has the characteristic features that distinguish it from Sequence G. However, the failure to obtain a reliable three-way Sequence × Format × Amount interaction ($F = 2.49$, $df = 2/96$, $p > .10$) allows us to retain the hypothesis stated in Experiment 1 that structural differences began quite early and did not change much throughout learning.

The middle row of Figure 3 shows the performance of the two instructional groups by problem type at each of three points in learning. Although Sequence F subjects performed better on F-type problems while Sequence G subjects performed better on Q- and U-type problems as was found in Experiment 1, the two-way Sequence × Type interaction failed to reach a statistically significant level ($F = 1.63$, $df = 4/384$, $p > .15$). An investigation of the marginally significant three-way interactions—Sequence × Type × Content ($F = 2.14$, $df = 8/768$, $p < .05$) and Sequence × Type × Format ($F = 2.40$, $df = 4/384$, $p < .05$)—indicates that the Sequence × Type interaction was strongest for B content and F format.

Since the Treatment × Posttest interaction was complicated by content and format and was not statistically significant in this case, there is little point in trying to locate where the Treatment × Posttest interaction began; however, there was no evidence that

the pattern of Treatment × Posttest interaction was influenced by the amount of instruction. The failure of the Sequence × Amount × Type interaction to reach statistical significance ($F = 1.17$, $df = 3/384$, $p > .20$) is consistent with the results of Experiment 1, indicating no structural change in how subjects in the two instructional groups encoded material over the course of learning.

An interesting fact, however, as observed with respect to problem format, is that Sequence G was superior to Sequence F on all problem types after Amount 1, but that they tended to even out after Amounts 2 and 3. A reliable Sequence × Amount interaction ($F = 3.30$, $df = 2/96$, $p < .05$) verifies this observation and permits the claim that adding the combinations sections (Amount 2) helped Sequence F subjects more that Sequence G subjects.

Finally, the performance of the two instructional groups by problem content at each of the three points in learning is shown in the bottom row of Figure 3. As in Experiment 1, a reliable Treatment × Posttest interaction suggests structural differences were obtained with Sequence F subjects superior on B-content and Sequence G superior to C- and J-content items ($F = 13.09$, $df = 2/192$, $p < .001$).

Again, the pattern of the Sequence × Amount interaction was that Sequence G outperformed Sequence F for all content areas after Amount 1 but evened out after Amounts 2 and 3. Again, the same general pattern of Treatment × Posttest interaction seems to have been present, at least in some form, at each of three points in learning. However, unlike Experiment 1, a reliable Sequence × Amount × Content interaction ($F = 2.62$, $df = 4/192$, $p < .05$) indicated that the structural differences changed as the amount of instruction was increased. Apparently, the presentation order used in Experiment 2, in which the more necessary content was not presented until Amount 2, helped focus on the disproportionate importance of specific content for incomplete structures of Sequence F subjects relative to Sequence G subjects.

Resilience of Structural Differences

Experiment 2 used the variable of time stress during testing (i.e., speed versus power) in another attempt to disrupt both the qualitative (or structural) and quantitative (or overall) features of the performance of subjects who had acquired different cognitive structures. Figure 4 shows the proportion correct response for the two instructional sequences by posttest item (i.e., Treatment × Posttest interaction) for each of two levels of time stress and for each of three posttest dimensions. The performance for the instructional groups under speed or time stress (right panel) and power or no time stress (left panel) conditions is given by problem format in the top row, by problem type in the middle row, and by problem content in the bottom row. Forcing subjects to answer within 90 seconds produced an overall decrement in performance ($F = 12.5$, $df = 1/96$, $p < .001$) manifested in the observation that the curves in the right panel are shifted down from those in the left panel. However, for each row, the same pattern of Treatment × Posttest interaction seems present both under time stress and no time stress. The failure to obtain a reliable Sequence × Format × Testing Condition ($F < 1.00$, $df = 1/96$, $p > .20$), Sequence × Type × Testing Condition ($F < 1.00$, $df = 4/348$, $p > .20$), or Sequence × Content × Testing Condition ($F < 1.00$, $df = 2/192$, $p > .20$) interaction confirms the lack of evidence that the 90-second limit had any effect on the structure or quality of learning outcomes.

Supplemental Studies

In addition to the two main experiments, two smaller, supplemental studies were conducted. The supplemental studies used the same instructional materials and procedures as the main studies but varied the testing situation in order to provide more information about the resilience and development of differences in learning outcomes.

Supplemental Study 1

One study focused on retention performance in an attempt to test the endurance

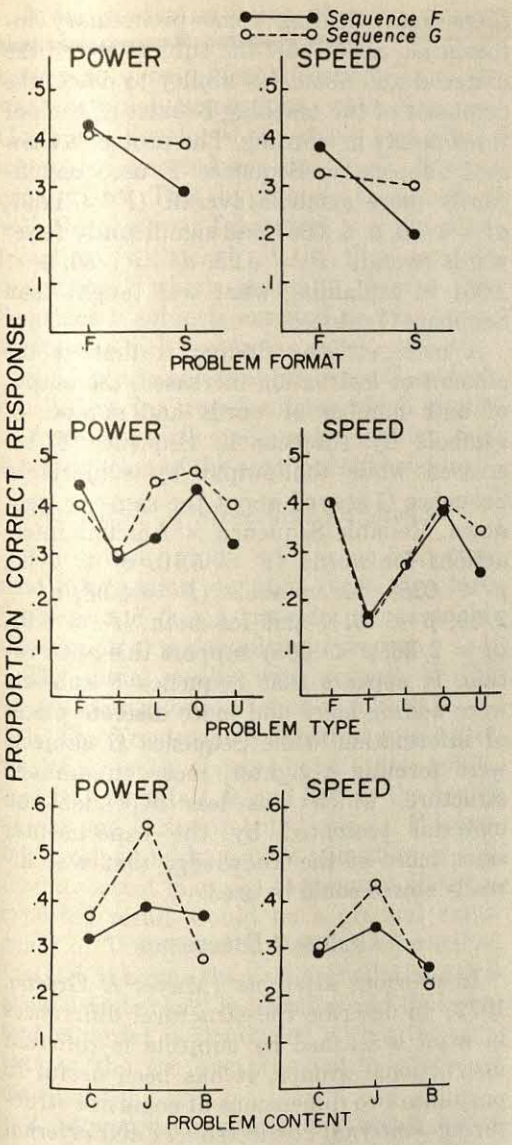


FIGURE 4. Treatment × Posttest interaction by testing condition for three posttest dimensions in Experiment 2. (Problem format: F = formula, S = story; problem type: F = familiar, T = transformed, L = Luchins, Q = question, U = unanswerable; problem content: C = combinations, J = joint probability, B = binomial probability.)

over time of the structural differences established by different instructional sequences.

Method

Nine subjects received the Sequence F, Amount 3, Order ABC booklet and nine other subjects received the Sequence G, Amount 3, Order ABC booklet. The 30-item transfer test was given—with no memory support and no time stress—two days later.

Results and Discussion

In partial replication of earlier findings, Sequence F subjects reached a higher proportion of correct response than Sequence G subjects on near transfer items such as F type (.59 versus .41), T type (.39 versus .33), F format (.54 versus .44), and B content (.44 versus .30), while Sequence G outperformed Sequence F on far transfer items such as U type (.44 versus .26), S format (.35 versus .30), C content (.44 versus .40), and J content (.43 versus .41).

As in the main studies, an analysis of variance revealed no significant difference between the groups in overall performance and found Treatment × Posttest interaction for type ($F = 3.58$, $df = 4/64$, $p < .025$) and Treatment × Posttest interaction for format ($F = 5.60$, $df = 1/16$, $p < .05$) to be at statistically reliable levels. The Sequence × Content interaction failed to reach significance ($F = 2.09$, $df = 2/30$, $p > .10$); however, this may be a reflection of the low number of subjects involved. A significant three-way Sequence × Type × Format interaction ($F = 5.41$, $df = 4/64$, $p < .001$) indicates that the Sequence × Type interaction was much stronger for F-format items than for S-format items. As with the memory-support and time-stress variables, there was no strong evidence that the retention interval altered the structure of learning outcomes.

Supplemental Study 2

The other supplemental study used a modified "method of reproduction" to assess a subject's judgment of what was supposed to be learned at each of three points in learning for the two instructional groups.

Method

Six subjects each were presented with either the Sequence F, Amount 1, Order ABC booklet; the Sequence F, Amount 2, Order ABC booklet; the Sequence F, Amount 3, Order ABC booklet; the Sequence G, Amount 1, Order ABC booklet; the Sequence G, Amount 2, Order ABC booklet; or the Sequence G, Amount 3, Order ABC booklet. All

TABLE 1

AVERAGE NUMBER OF ELEMENTS IN REPRODUCTION FOR TWO INSTRUCTIONAL GROUPS AND THREE AMOUNTS OF INSTRUCTION—SUPPLEMENTAL STUDY 2

Instructional sequence	Amount of instruction			
	1	2	3	Average
Average no. words				
F	110	152	243	168
G	259	278	206	247
Average	184	215	225	
Average no. symbols				
F	95	187	190	157
G	111	107	88	102
Average	103	147	139	
Average no. words and symbols				
F	205	339	433	326
G	370	385	293	350
Average	287	362	363	

Note. Abbreviations: F = formula and G = general concepts.

subjects were then given an immediate retention test and questionnaire. The test was to write down, with no memory support or time stress, what had been in the teaching program, as if the subject were trying to teach it to another subject coming in for the next session. The traditional method of reproduction was modified in that the subject was encouraged to "make sense" out of the material as if he were explaining what he had just learned to someone who did not already know it rather than to reproduce the material word for word as given.

Results and Discussion

The protocols from each subject were coded for number of words and number of symbols, with each operator, each number, and each formal notation character counting as a symbol (e.g., N! counts as 2 symbols; 15/2 counts as 3 symbols). Table 1 shows the average number of words, the average number of symbols, and the average number of both words and symbols given at each point in learning by subjects in the two instructional groups. Separate analyses of variance were performed on these data yielding some preliminary information about how the subject stored the material and about his ability to detect the emphasis of the teaching booklet at each of three points in learning. The protocols show that subjects in Sequence F used significantly more symbols overall ($F = 12.97$, $df = 1/30$, $p < .005$) and significantly fewer words overall ($F = 9.25$, $df = 1/30$, $p < .005$) in explaining what was taught than Sequence G subjects.

A more striking finding is that as the amount of instruction increased, the output of both number of words and number of symbols by subjects in Sequence F increased while the output by subjects in Sequence G stayed about the same or went down. Reliable Sequence × Amount interactions for words ($F = 5.10$, $df = 2/30$, $p < .025$), for symbols ($F = 5.63$, $df = 2/30$, $p < .01$), and for both ($F = 6.91$, $df = 2/30$, $p < .005$) support this observation. It appears that Sequence F subjects were adding more and more discrete pieces of information while Sequence G subjects were forming a tighter, more streamlined structure, which was less dependent on material presented by the experimenter since more of the knowledge that was already stored could be used.

GENERAL DISCUSSION

In previous attempts (Mayer & Greeno, 1972) to describe the structural differences in what is learned by subjects in different instructional groups, it has been useful to postulate two dimensions of cognitive structuring—*internal connectedness* and *external connectedness*. Internal connectedness refers to the degree to which the variables of the formula are related to one another in the subject's cognitive structure, for example, P is related to R by exponentiation. External connectedness refers to the degree to which variables of the formula are related to knowledge already existing in the subject's cognitive structuring, for example, P is related to past experience with probability such as weather forecasts (20% chance of rain), dice (each number has ⅙ probability), batting average (.333 means the batter gets a hit ⅓ of his times at bat).

In the present situation, there is evidence that Sequence F subjects developed structures with strong internal connectedness but weak external connections, while Sequence G subjects developed structures with strong external connectedness but weak internal connections. This hypothesis is consistent with the observed Treatment × Posttest interaction (with each of the three posttest dimensions) which generally showed that Sequence F subjects excelled on near transfer items which require an exact application of the formula and that Sequence G subjects excelled on far transfer items which require a more sophisticated understanding and interpretation of the component variables.

Development of Structural Differences

The present study adds some important new information on how these structures develop and helps provide a reasonable description of the acquisition process.

The main results with respect to this "acquisition question" were as follows:

1. The Treatment × Posttest interaction was consistently observed at all three points, that is, Amounts 1, 2, and 3, over the course of learning (Experiments 1 and 2). If learning involved adding more and more of different content material to memory, then the expected result would be a gradual emergence of Treatment × Posttest interaction. Instead, it seems than an assimilative set is evoked quite early in learning and that content material is structured within the context of the set over the entire course of learning.

2. Sequence G subjects performed better overall relative to Sequence F subjects when specifically needed content material, that is, the combinations lessons, was lacking (Experiment 2). One interpretation is that the richer assimilative set of Sequence G subjects, that is, the rich body of existing relevant knowledge, helped create original problem solutions while the narrower assimilative set of Sequence F subjects made problem solving more dependent on specific content material.

3. As the amount of instruction was increased, Sequence F subjects produced longer and longer reproduction protocols, while the protocols of Sequence G subjects remained about the same or decreased (Supplemental Study 2). Again, these results seem most consistent with the idea that Sequence F subjects were adding material to fairly narrow assimilative sets while Sequence G subjects were integrating and streamlining material to fit within a rich bank of existing knowledge.

The results of the present study seem consistent with the notion that different receptive sets are activated early in learning. Furthermore, the differences in how the material is encoded are established early and remain constant throughout learning. Greeno (1972) has provided a framework for analyzing problem-solving behavior that relies on a distinction between two kinds of knowledge stored in semantic memory—"propositional knowledge" and "algorithmic knowledge." Propositional knowledge refers to relational and conceptual information such as hierarchies of classes and subsets (collies are dogs), properties of classes of things (dogs have tails), and facts (April 17–23 is National Dog Week). Algorithmic knowledge refers to rules or operations such as the procedures followed in doing long division.

In the present example, it seems that the organization and emphasis of Sequence F encourages the use of a narrow, assimilative set concerned with mathematical operations and calculations—what could be called algorithmic knowledge—and that Sequence G encourages the activation of a broader, more integrative set made up of the subject's general experience—what could be called propositional knowledge. Since what is learned is the product of both the presented material and the assimilative set the subject uses to encode it, different learning outcomes are possible. What these results seem to show is that the algorithmic kind of set is activated quite early in learning for Sequence F and that the propositional kind of set is activated quite early in learning for Sequence G; further, there is support for the idea that the structural differences observed at the end of learning have their roots in this embedding of material into two different kinds of assimilative sets or of knowl-

edge which begins early and continues throughout learning.

Resilience of Structural Differences

The main results with respect to the "resilience question" were as follows: (*a*) Memory-support or open-book testing had a mildly positive overall effect on problem-solving performance, but there was no evidence of any change in the pattern of Treatment × Posttest interaction (Experiment 1). (*b*) Time stress, the speed condition, had a strong overall inhibitory effect on problem-solving performance, but there was, again, no evidence of any effect on the pattern of Treatment × Posttest interaction (Experiment 2). (*c*) A two-day retention interval also failed to produce disruptions of the usual patterns of Treatment × Posttest interaction (Supplemental Study 1). Thus, all attempts to alter the structure or quality of problem-solving performance failed. Apparently, once the problem-solving rule is encoded, testing conditions can influence the absolute level of performance but with negligible effects on the structure of problem-solving performance.

Relation to Discovery Learning and Creative Problem Solving

These results also give some hints about the prerequisites for discovery learning and creative problem solving. For the kind of teaching whose goal is creative problem solving, such as displayed by Sequence G, Amount 1 subjects in Experiment 2 (i.e., inventing solutions with very little specific content material having been given), it is clear that a substantial bank of what has been called "propositional knowledge" must be available. For subjects who do not have a well integrated set of general experiences in the required area (e.g., in this case, in probability of events), it seems that attempts to achieve learning outcomes that can support creative problem solving will fail. This is so because the kind of learning outcome that supports creative problem solving (e.g., as displayed by Sequence G subjects) is acquired by embedding the problem-solving rule in a bank of the appropriate propositional knowledge. Instead of demanding that subjects who lack the appropriate prerequisite knowledge learn a given problem solution rule, thereby insuring either no learning or a very specific kind of encoding, it seems a better educational practice to first provide subjects with the necessary prerequisite concepts. Previous findings (Egan & Greeno, 1973) that show individual differences in specific prerequisite knowledge to be far more important for "discovery" learning methods than for "rule" learning methods support this argument.

In short, the results seem to indicate that for those kinds of learning whose goal is a quick, efficient ability to perform a given set of operations (e.g., arithmetic operations, etc.), only a narrow set of existing knowledge is necessary; but for learning that supports creative problem solving (e.g., reconstructing a procedure for new problems), the need to make sure the subject possesses the prerequisite knowledge is essential. These results suggest that discovery teaching to subjects who lack the appropriate propositional knowledge will not result in discovery learning or creative problem solving and may result in no learning at all.

REFERENCES

Ausubel, D. P. *Educational psychology: A cognitive approach.* New York: Holt, Rinehart & Winston, 1968.

Egan, D. E., & Greeno, J. G. Acquiring cognitive structure by discovery and rule learning. *Journal of Educational Psychology*, 1973, **64**, 85–97.

Greeno, J. G. *The structure of memory and the process of solving problems.* (Tech. Rep. 37) Ann Arbor: University of Michigan, Human Performance Center, 1972.

Mayer, R. E. Acquisition and resilience under test stress of structurally different problem solving procedures. (Doctoral dissertation, University of Michigan) Ann Arbor, Mich.: University Microfilms, 1973.

Mayer, R. E., & Greeno, J. G. Structural differences between learning outcomes produced by different instructional methods. *Journal of Educational Psychology*, 1972, **63**, 165–173.

Roughead, W. G., & Scandura, J. M. "What is learned" in mathematical discovery. *Journal of Educational Psychology*, 1968, **59**, 283–289.

(Received September 28, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 657-662

LONGITUDINAL CHANGES IN INTELLECTUAL FUNCTIONING IN PRESCHOOL CHILDREN: SCHOOLING- AND AGE-RELATED EFFECTS

L. R. GOULET,[1] KERRY G. WILLIAMS, AND CARL M. HAY

University of Illinois at Urbana-Champaign

The mental age (MA) and IQ of three samples of nursery school children varying in chronological age (CA) were ascertained at two points in the school year. The testing points were chosen to conform to a design which permitted the non-CA-related influences of schooling to be estimated. The results revealed that both CA and amount of time in school were positively related to MA. A similar relation between IQ and length of schooling was also found. The implications of these results were discussed in terms of both the direct and indirect influences of schooling on intellectual functioning and in terms of potential sampling biases inherent in the norming procedures for available standardized tests.

One of the objectives of developmental investigations is the identification of the components which contribute to age-related changes in performance. Schaie (1965), for example, has suggested that individuals can be described according to three such components: chronological age (CA), cohort (birth date), and time of measurement. Furthermore, he suggested that performance differences between populations varying in chronological age represent the composite effects of these three components rather than of age changes alone and that conventional research designs used in developmental research confound these components.

For example, age differences found using cross-sectional research designs (where samples of subjects varying in chronological age are tested at the same point in time) may represent age changes, cohort changes, or their joint influences. Similarly, age differences in performance resulting from the use of a longitudinal research design (where subjects from the same cohort are tested at different points in time) may be attributed to age changes, to time of measurement differences, or to their joint influences. Such inferences point to the limited usefulness of using simple longitudinal or cross-sectional designs in descriptive developmental research and also suggest a possible explanation for the disparities in the nature of developmental functions obtained using these data collection strategies.

The present study had three purposes: (*a*) to compare longitudinal and cross-sectional age gradients in intellectual functioning in young children; (*b*) to identify age-related and time-related (specifically nursery school experiences) changes on such intellectual functioning; and (*c*) to examine the stability of the estimate of IQ over periods of time when the population of interest was enrolled in nursery school.

Available data for school-age children suggest that longitudinal age gradients in intellectual functioning provide estimates of change which exceed the estimates provided by cross-sectional age gradients (e.g., Baltes & Reinert, 1969; Schaie, 1972). Such a disparity may be expected, since longitudinal measurement, by definition, provides estimates of change which confound age-related and school-related influences on performance. In other words, over the period of a school year, the child is not only profiting

[1] Requests for reprints should be sent to L. R. Goulet, Department of Educational Psychology, University of Illinois, Urbana, Illinois 61801.

from the influences of school, he is "aging" or developing. In contrast, within-grade cross-sectional comparisons do not incorporate such confounding, since performance is assessed for all samples at the same point in time (thereby controlling for the amount of time spent in school). The present data provide evidence regarding the generalizability of this phenomenon for nursery-school-age children.

Correlated with the first purpose is the present concern with identifying the variance associated with age and school experience during the period of time that the child is enrolled in nursery school. That is, even though age and amount of schooling are inextricably correlated for individual children, it is possible to obtain independent estimates of the influences of schooling. This is possible since children within a grade vary in the age of school entry. Thus, it is possible to compare the performance of a sample of subjects at a point early in the school year with that of a matched CA sample tested at a later point in the school year. The available data permitting such comparisons suggest that the influences of schooling equal or exceed those related to CA for samples enrolled in the first grade (Schaie, 1972). Similar conclusions are possible at least through the fourth grade (Baltes, Baltes, & Reinert, 1970; Baltes & Reinert, 1969). Again, however, no data are available for younger children.

Finally, the present study provides data concerning the sensitivity of the IQ measure to the influences of schooling. The Peabody Picture Vocabulary Test (Dunn, 1965), which was used for present purposes, was standardized using cross-sectional sampling procedures during the months of April–June. None of the children younger than six years old in the normative sample were enrolled in school. The present testing took place at two points in the school year, over which three samples of children varying in CA were each tested twice. Thus, at the second testing period, the children were both older and had more school experience. It follows, therefore, that the IQ scores for the second time period should exceed those obtained for the first time period even though the transformation from mental age (MA) to IQ controls for influences related to CA.

Method

Design

The experimental design was a 3 × 2 (Chronological Age × Time of Testing) factorial. All subjects were tested twice, once in the fall of the school year (October–November) and once in the spring (March–April). All subjects were administered Form A of the Peabody Picture Vocabulary Test in the fall, whereas half of the subjects received Form A and the remainder received Form B at the spring testing. Approximately equal numbers of boys and girls were included in each CA group.

Subjects and Procedure

The subjects were 63 preschool children (31 boys, 32 girls) enrolled in one of three preschools in Champaign-Urbana, Illinois. The subjects were randomly chosen from those available within the three age groups. The mean CA for the young, middle, and old groups at the two times of testing are contained in Table 1.

The standard procedures for administration of the Peabody Picture Vocabulary Test were followed with the exception that testing began at Item 1 of the test. The MA and IQ scores were taken from tables available in the test manual (Dunn, 1965). Testing was conducted in a test trailer or private room at each of the nursery schools. Retesting in the spring occurred in the approximate order of testing used in the fall. This ensured that the mean age of each group increased by five months and that the distribution of ages within groups remained stable from fall to spring.

Results

Longitudinal and Cross-Sectional Age Gradients

The present data provide two sets of cross-sectional comparisons, that is, one each for the sets of data from each testing period. The data from the first testing period are most appropriate for present purposes and are presented in Figure 1 along with the longitudinal gradient for the same age range estimated from the present data. The longitudinal gradient was constructed by assessing the change in MA from the first to the second testing period for the CA range from 46 months to 52 months and similarly for the CA range from 52 to 57 months, and adding these changes to the

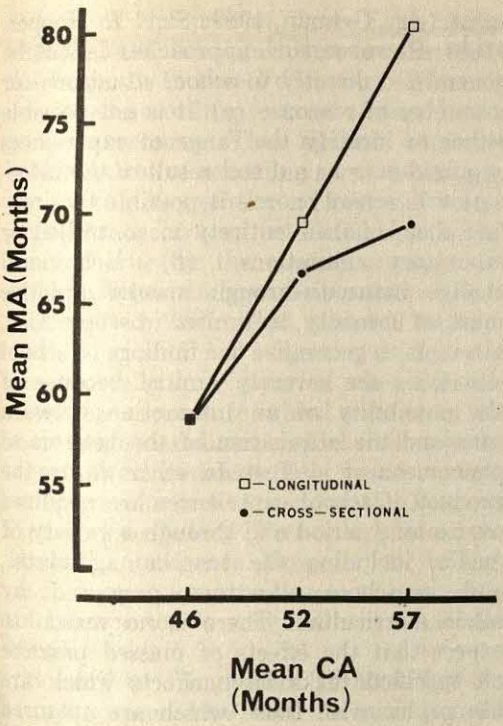


FIGURE 1. Cross-sectional and longitudinal age gradients for nursery school children. (Abbreviations: CA = chronological age and MA = mental age.)

known base for the youngest sample (Schaie & Strother, 1968). As is apparent from Figure 1, the longitudinal age gradient provides estimates of age-related change which exceed those of the cross-sectional age gradient and suggests the existence of facilitative effects related to schooling. In addition, both gradients reveal a positive relationship between CA and performance.

Table 1 provides summary data related to the chronological age, mental age, and IQ for the samples of subjects at the two time periods. Preliminary statistical analyses for each of the measures revealed no differences in performance attributable either to sex of the subject or to the form of the test which was administered to the subjects. Therefore, the data were pooled. The sets of data relating to MA and IQ were each subjected to a 3 × 2 (Chronological Age × Time of Testing) analysis of variance. Time of testing was a repeated measure.

The analysis for the data reflecting changes in MA revealed statistical significance both for the main effects of CA ($F = 6.25$, $df = 2/60$, $p < .003$) and time of testing ($F = 66.64$, $df = 1/60$, $p < .001$). It is apparent, however, that the influences of time of testing represent sources of variance associated both with age change and with factors associated with school experience. Fortunately, the design permits an evaluation of performance differences between the two times of testing for matched CA samples. Two such contrasts are possible, that is, at CA of 52 months and CA of 57 months. The pooled within-cell error variance (Winer, 1962) was used as the error estimate for the contrasts. Only the contrast for the older samples reached statistical significance ($p < .05$), even though both comparisons yielded data in the predicted direction.

The final statistical analysis involved a comparison of IQ scores. The longitudinal contrasts provide an alternate way of evaluating the influences of school experience in that the age-related changes are controlled when the mental age scores are transformed to IQ. The results of this analysis revealed statistical significance for the main effect of time of testing ($F = 21.98$, $df = 1/60$, $p < .001$). The absence of statistical significance for the main effect of CA and the interaction (both Fs < 1) suggested equivalence of the samples in measured IQ, and statistical contrasts between matched CA samples again revealed statistical significance for the comparison involving the older samples

TABLE 1

SUMMARY DATA FOR YOUNG, MIDDLE, AND OLD SAMPLES AT TWO TIMES OF TESTING

Sample	Measures					
	CA		MA		IQ	
	T_1	T_2	T_1	T_2	T_1	T_2
Young	46.5	51.8	58.1	69.3	110.1	115.2
Middle	52.2	57.1	66.9	78.1	111.7	117.8
Old	56.8	62.2	69.3	83.4	110.6	118.9

Note. Abbreviations: CA = chronological age, MA = mental age, and T = time.

($p < .05$) but not for the younger ones. Such a conclusion is also suggested from the longitudinal differences which reflect a mean IQ difference of 5.1 points between times of testing for the youngest sample, with this difference increasing to 8.3 for the oldest sample (see Table 1).

Discussion

The present results suggesting that schooling has effects on performance independent of those related to CA generally confirm the findings of studies (Baltes et al., 1970; Baltes & Reinert, 1969; Schaie, 1972) with older children. In fact, the examination of the MA scores for the three samples reveals that the longitudinal changes over the two testing periods exceed by 100% or more the relatively substantial changes which are attributable to CA alone (Dunn, 1965). Again, these results confirm the findings of others (e.g., Schaie, 1972).

It is perhaps most important to discuss the implications of the present findings for educational and developmental research and measurement, and in the assessment of the effects of educational experiences.

The primary issue to be considered here concerns the assessment of the effects of educational intervention (used in the broad sense) on performance over the period of a school year. In this context, a severe and limiting source of confounding is potentially present; that is, the influences of schooling are not discernible from other CA-related influences on performances. That is not to say that the impact of or effects of exposure to the school curriculum can be considered to be independent of behavioral development. Rather, school learning must be considered to be one of the components in the developmental process. If for no other reason than this, the effects of school experience must be estimated.

The most obvious way in which to provide estimates of the effects of experience on performance unbiased by behavioral development is to simulate or "accelerate" the process through the provision of massed training or practice. Such an experimental strategy is used very often in contemporary studies concerned with cognitive development (e.g., Gelman, 1969; Sigel & Hooper, 1968). However, such approaches cannot be generalized directly to school situations for a number of reasons: (*a*) It is not possible either to identify the range of experiences acquired in or as a direct result of the interaction in school; nor is it possible to simulate them in their entirety in controlled or laboratory situations. (*b*) Behavioral change induced through massed practice must, of necessity, be limited in scope. Also, attempts to generalize the findings to school situations are severely limited because of the possibility of an interaction between time and the acquisition of the behavioral phenomena of interest. In other words, the product of school experiences are acquired over a long period and through a variety of media, including the teacher, age-mates, and nonschool situations prompted by school curriculum. There is no reason to expect that the effects of massed practice on specified tasks have effects which are isomorphic with those which are acquired as a result of schooling over the school year.

There is a second way to provide direct estimates of the effects of school experience which are unbiased by independent time- or age-related components of behavioral change. In the most simple case, the procedure would involve the comparison of two groups of children across time (e.g., the school year) under conditions in which both groups were eligible for acceptance into school but one of the two groups was enrolled in school and one was not. However, it is extremely difficult to find a sample of children who are of school age but who have not been enrolled in school. And, even if such a sample were available in the general population, it would be impossible to match them with children who were enrolled. The very conditions which precipitated the lack of enrollment would bias the sample. As is apparent, the present study utilized a research strategy which capitilizes on the latter method while avoiding the potential sources of confounding when it is used.

The present data also relate to procedures of test development, standardization, and norming. There are two primary impli-

cations of the present data. First, it has already been mentioned that the norms for the Peabody Picture Vocabulary Test were developed according to cross-sectional sampling procedures. This method implicitly holds amount of schooling constant for subjects within a grade or, of course, for subjects who are not as yet enrolled in school (as was true for the normative sample of children below the age of six years). It is to be expected, then, that the IQ estimate would increase if the subjects were followed longitudinally over a period when the subjects were exposed to school or to other environmental conditions which influence intellectual achievement. This would be especially true for a test such as the Peabody Picture Vocabulary Test which measures "verbal" intelligence. However, for within-grade comparisons, the degree of bias would not be expected to vary with the CA of the sample.

Second, it is important to note that the sampling procedure used in norming the Peabody Picture Vocabulary Test entertains a quite different type of bias for preschool children than for populations of children enrolled in school. Again it must be noted that the procedure used for norming the Peabody Picture Vocabulary Test involved a decidedly biased sample. That is, all children above the ages of six and a half years in the normative sample were enrolled in school and, because of the testing period (April–June), had been enrolled for most or all of a complete school year. Unfortunately, birth dates and birth rates in the population do not distribute themselves in this manner but rather are distributed throughout the calendar year. This means, of course, that the test scores (MA or IQ) for school-age children who are tested in the months from July to March would be compared to that of a normative sample which enjoyed from seven months to a minimum of one month more schooling, respectively. Such contrasts suffer from a negative bias which would decrease in magnitude as the data of testing approaches that for the normative sample. Such a bias may also exist for other tests, but the degree and direction of bias would be expected to vary with the type of sampling procedure used in the test standardization and norming, in addition to the nature of the relationship between amount of school experience and performance.

The present results also suggest the utility of norming tests using samples selected and tested at different times within a year. This suggestion has had its proponents (e.g., Baltes & Reinert, 1969; Lodge, 1938) but has never been formally implemented.

Some final comments concerning the influences of schooling are warranted. First, there is no intent to imply that the results attributed to the influences of school experience in the present study are directly or exclusively attributable to the "in-classroom" experiences of the children. Rather, such influences may take many forms, ranging from the effects of the different forms of social interactions, environmental contexts, and parental or peer demands which confront the children while they are enrolled in school. It should also be emphasized that the character of extraschool experiences may vary considerably for children who are and who are not enrolled in school during the same CA period. In the present study, for example, the three CA samples entered school at different CAs and, therefore, spent varying amounts of time at home. The influences of such factors cannot be discounted in the present results. Last, the possibility exists that schooling may have indirect effects on performance even before the child enters school. As an example, parents may react to the child's impending entry into school by providing certain forms of educational experiences (e.g., reading, counting, etc.) at home to "prepare" the child for school. There is evidence to support such a contention (Hay & Goulet, 1973), although such factors were not of interest in the present study.

REFERENCES

Baltes, P. B., Baltes, M. M., & Reinert, G. The relationship between time of measurement and age in cognitive development of children: An application of cross-sectional sequences. *Human Development*, 1970, **13**, 258–268.

Baltes, P. B., & Reinert, G. Cohort effects in cog-

nitive development of children as revealed by cross-sectional sequences. *Developmental Psychology*, 1969, **1**, 169–177.

Dunn, L. M. *Peabody Picture Vocabulary Test: Manual*. Circle Pines, Minn.: American Guidance Service, 1965.

Gelman, R. Conservation acquisition: A problem of learning to attend to relevant attributes. *Journal of Experimental Child Psychology*, 1969, **7**, 167–187.

Hay, C. M., & Goulet, L. R. Age, sex, and school entry as cues for training and sequences of achievement in young children. (Research Rep. No. 103) Urbana: University of Illinois, Department of Educational Psychology, 1973.

Lodge, T. Variation in Stanford-Binet IQs of preschool children according to the months in which examinations were given. *Journal of Psychology*, 1938, **6**, 385–395.

Schaie, K. W. A general model for the study of developmental problems. *Psychological Bulletin*, 1965, **64**, 92–107.

Schaie, K. W. Limitations on the generalizability of growth curves of intelligence. *Human Development*, 1972, **15**, 141–152.

Schaie, K. R., & Strother, C. R. A cross-sectional study of age changes in cognitive behavior. *Psychological Bulletin*, 1968, **70**, 671–680.

Sigel, I. E., & Hooper, F. H. *Logical thinking in children: Research based on Piaget's theory*. New York: Holt, Rinehart & Winston, 1968.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received November 8, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 663-668

DOLL PREFERENCES:

AN INDEX OF RACIAL ATTITUDES?[1]

PHYLLIS A. KATZ[2] AND SUE ROSENBERG ZALK

Graduate School and University Center of the City University of New York

A doll choice task was administered to 192 nursery and kindergarten children by same- and other-race examiners. Half of the children were white and half were black. In contrast to earlier studies, male and female dolls were presented which differed only in skin color, not hair or eye color. The strong preference for white dolls found by previous investigators was not obtained. Young children exhibited a slight preference for other-race dolls, although gender cues were more significant determinants of choice behavior than were skin color cues. Children's responses were, in part, a function of the tester's race. Stronger preferences for same-race dolls were exhibited in the presence of a same-race examiner.

The acquisition of racial attitudes in young children has frequently been studied. Contrary to the opinion of the lay public, the most consistent finding has been that preschool children are very much aware of racial cues. By the age of four, most children can correctly identify skin color. Moreover, there appears to be a considerable degree of evaluative content associated with racial cues (e.g., Asher & Allen, 1969; Clark & Clark, 1947; Greenwald & Oppenheim, 1968; Porter, 1971).

The assessment of young children's racial attitudes typically utilizes a task involving doll choices. The child is presented with several dolls and is asked to choose the doll that goes with either a positive or negative characteristic (i.e., which is the good doll). Clark and Clark (1947) originally demonstrated that preschool children (both black and white) tend to associate positive attributes to the white doll and negative ones to the black one. This finding has been replicated many times. Both the consistency of these results and the seemingly straightforward nature of the doll task undoubtedly accounts for its continued popularity as an index of young children's racial attitudes and perceptions.

Despite its general acceptance, however, the typical administration of a doll preference task raises a number of interpretive and methodological problems. The first one concerns the psychometric characteristics of the instrument itself. Neither reliability nor validity data has ever been presented. It has never been shown that children's preferences are consistent over time nor that a child's response on a doll task is related to future attitudes. Moreover, the only study attempting to relate doll preferences to intergroup behavior (friendship choices in nursery school) found no relationship (Hraba & Grant, 1971). Thus, the validity of this measure remains to be demonstrated. The assumption made by previous investigators that a prejudiced child will remain so clearly needs substantiation.

[1] This research was supported by Grants HD 04821-01 and HD 04821-02, National Institute of Child Health and Human Development, Phyllis A. Katz, principal investigator. The write-up was supported by Grant OCD-CB-151 from the Office of Child Development.

The authors are most grateful to the principals of Public School 26, The Bronx YMHA Nursery School, and Grant Houses Day Care Center for their generous cooperation in providing subjects for this study. Thanks are also extended to Judith Kroll, Joyce Johnson, Karen Codrington, Carol Seavey, and Marilyn Sohn for their assistance in data collection and analysis.

[2] Requests for reprints should be directed to Phyllis Katz, Graduate Center, City University of New York, 33 West 42nd Street, New York, New York 10036.

A second problem has to do with interpreting the relative importance of racial cues in the child's judgment. The impression earlier studies impart is that skin color cues are very significant indicators of a young child's perception of people. A recurrent methodological difficulty, however, has been the consistent confounding of skin color with other cues. White dolls typically have blue eyes and blonde hair, whereas black dolls have brown eyes and brown hair. The possibility exists, therefore, that children may have based their choices on cues other than skin color. An additional difficulty is that only female dolls are ordinarily displayed, thus eliciting preferences that may not be generalizable to males. Moreover, when a choice is limited to only race cues, the strength of those cues relative to other person characteristics (such as gender) may be seriously overestimated.

A final problem with most earlier studies concerns the type of examiner employed. Recent evidence suggests that the race of the examiner may be an important variable influencing children's performance in a wide variety of contexts (Bucky & Banta, 1972; Katz, 1973; Katz, Johnson, & Parker, 1970; Sattler, 1970). Nevertheless, most earlier studies with preschool children have not varied the race of the examiner.

The present study attempted to clarify some of these methodological ambiguities in children's doll preference behavior. In view of the previously raised considerations, dolls were employed which varied only in skin color, with hair and eye color constant. Gender cues were provided in order to assess the relative strength of racial and sexual preference. An inquiry procedure was included in which reasons for the child's choice were elicited. The purpose of this latter task was to ascertain whether children's verbal behavior was related to doll choices. Finally, half of the subjects in each group were tested by a black and half by a white tester in order to explore the effects of race of examiner.

Method

Subjects

A sample of 192 subjects was employed, equally divided according to age, race, and sex. The mean chronological ages of the younger and older groups were 3 years 11 months and 5 years 2 months, respectively, with corresponding standard deviations of 5 and 6 months.

Subjects were drawn from several sources. Older children were randomly selected from kindergarten classes of an integrated public elementary school located in a lower to lower-middle income area of New York City. The younger children were selected from two nursery schools in the same geographical area. At the first nursery school (a church-associated one), 75% of the students were white and 25% were black. The second nursery sample was drawn from a city-run nursery school which was predominantly black and which was located in a neighborhood that was of a somewhat lower socioeconomic level. Analyses of the responses of the black children, however, revealed no differences between the two nursery settings so that they were pooled in subsequent analyses.

Doll Preference Task

Two black and two white dolls, one of each sex, were introduced to the children. The dolls were approximately 10 inches tall, made of rubber, and were all cut from the same mold. They all had brown eyes. The same dark brown wigs were used on all dolls. The hair was straight and coarse textured and was judged by adults (both black and white) to be appropriate for either race. Facial features were constant across skin color, and noses and mouths were judged by adults to be ambiguous enough to be either European or African toddlers. The male dolls had short hair and were dressed in a white shirt and red pants; the female dolls had longer hair and wore white shirts and red skirts. The children were asked to choose (*a*) the doll they like best; (*b*) the doll they didn't like as much as the others; (*c*) the good doll; (*d*) the bad doll; (*e*) the doll that was a nice color; (*f*) the doll that was not a nice color; and (*g*) which doll they would prefer to take home.

In addition, three identification questions were employed in which the child was asked to point to (*a*) the doll that looked most like him or her; (*b*) the one that looked like a white child; (*c*) and the one that looked like a black child. Following the questions, an inquiry procedure was employed in which the child was encouraged to give reasons for his choice, for example, Why is that one a good doll? The frequencies of doll choices made to each question were tabulated for each group. In addition, a composite race preference and sex preference score was obtained for each subject for the positive and negative items. An accuracy of identification score was also obtained. Finally, verbal responses were categorized according to the attributes spontaneously mentioned in the inquiry portion.

General Procedure

Each subject was individually tested in a room at their respective school. Within each age and racial group, half the subjects were tested by a white examiner and half by a black examiner. The

examiners were four females, in their middle twenties, who had had previous experience working with young children.

Results

There are a number of ways of analyzing the choices that children made with regard to the four doll choices. In earlier studies, frequencies to each question have been tabulated and nonparametric tests have been conducted. It was felt that a more efficient way of handling the data in the present study was to combine responses to the various questions and to assign each child a racial and sexual preference score. The racial preference score was determined by the number of times that the child chose a same-race doll for positive items and an other-race doll for negative items. Similarly, a sexual preference score for each child was determined on the basis of the number of times the child chose a same-sexed doll for positive characteristics and other-sexed doll for negative characteristics.

Racial Preference

The average racial preference scores for the positive and negative items are contained in Table 1. Since there was an unequal number of positive and negative items, each average score was further divided by the number of items within each subtest in order to make them comparable. Thus, the possible range was from 0 to 1. Lower scores indicate choices of other-race dolls for positive items and same-race dolls for negative attributes. High scores indicate the opposite pattern, that is, same-race choice for positive items and other-race choice for negative items. A score of .5 indicates a chance distribution of choices. It can be seen in Table 1 that, for the most part, the children's scores did not differ much from chance expectancy. This trend was in contrast to earlier results obtained by Clark and Clark (1947) and others, which showed that both black and white preschool children have strong preferences for white dolls. If the results presented in Table 1 conformed to earlier results, the white children should have averaged about .7, whereas the black subjects should have received scores of approximately .3 (indicating other-race choices).

TABLE 1
Average Racial Preference Scores on Doll Choice Task

Age/race of subject	Race of examiner	Type of item		
		Positive	Negative	Combined
Nursery				
White	same	.44	.52	.48
	other	.38	.40	.39
Black	same	.29	.44	.36
	other	.34	.45	.40
Kindergarten				
White	same	.63	.65	.64
	other	.40	.36	.38
Black	same	.55	.52	.54
	other	.44	.50	.47

A four-way analysis of variance conducted on the racial preference scores (Age × Race of Examiner × Race of Subject × Type of Item) revealed the following effects to be significant: Age ($F = 3.90$, $df = 1/184$); Type of Item ($F = 4.58$, $df = 1/184$); Race of Examiner × Race of Subject ($F = 3.92$, $df = 1/184$); and Age × Type of Item ($F = 4.18$, $df = 1/184$). The age effect indicates that nursery school children expressed less prejudice than kindergarten children, a mean of .41 as compared to .51. Actually, the overall average for the older children, as noted above, did not differ from chance expectation, and the lower score obtained by the younger children indicates a slight preference for other-race dolls. Doll choices are influenced by whether children are being tested by a same- or cross-race examiner. The significant Race of Examiner × Race of Subject interaction reflects the finding that children are more prone to select dolls of their own skin color for positive items and other-race dolls for negative attributes when tested by an examiner of the same race. This trend appears more pronounced in white than in black children. The performance of the white kindergarten children is particularly interesting since they show a preference for white dolls with the white examiner but a preference for black dolls with the black examiner. The type of item effect reflects the finding that children had a greater tendency to express attitudes of greater

TABLE 2

AVERAGE SEX PREFERENCE SCORES ON DOLL CHOICE TASK

Age/sex of subject	Type of item		
	Positive	Negative	Combined
Nursery			
Male	.48	.50	.49
Female	.64	.50	.57
Kindergarten			
Male	.50	.48	.49
Female	.79	.58	.68

tolerance with regard to positive rather than negative attributes. This tendency was more pronounced for the nursery group, as indicated by the Age × Type of Item interaction.

Sexual Preference

Since the children could choose the dolls on the basis of either skin color or gender cues, an analysis of variance was also conducted on the sex preference scores of the children. These scores are contained in Table 2.

Sex preference scores were obtained in the same way as the race preference scores. A mean of .50 indicates no preference, a higher score indicates preference for same-sexed dolls on positive items and other-sexed dolls on negative ones, whereas a lower score reveals the reverse pattern.

The analysis revealed that the main effects of sex of subject ($F = 12.03$, $df = 1/176$), type of item ($F = 10.68$, $df = 1/176$), and the Sex of Subject × Type of Item interaction ($F = 9.40$, $df = 1/176$) to be significant. As can be seen in Table 2, the significant results indicate that girls had much stronger preferences for same-sexed dolls ($M = .63$) than did boys ($M = .49$), particularly with regard to positive attributes. None of the other main or interaction effects were significant in the analysis.

A comparison of Tables 1 and 2 is interesting in that it indicates that, at least for girls, gender appears to be a much stronger cue for doll choice than skin color.

Racial Identification

The average racial identification scores are presented in Table 3. Since there were three racial identification questions, the possible range of scores on this measure was from 0 to 3, with higher scores associated with greater accuracy. As can be seen in Table 3, most of the children had very high scores on this measure. An analysis of variance revealed a significant age effect ($F = 11.62$, $df = 1/176$) with nursery subjects obtaining a mean of 2.43 as contrasted with 2.73 for the older children. In addition, the interaction of Age × Race of Subject × Sex was significant at the .05 level ($F = 6.51$, $df = 1/176$). As can be seen in Table 3, this interaction appears to be accounted for primarily by the relatively low scores of the black male nursery school children. Interestingly, this is the same group which shows a stronger tendency to select white dolls for positive items. This accounts for the only significant intermeasure correlation obtained, namely, a correlation of .27 between racial identification and racial prejudice scores.

Relation of Verbal and Choice Behavior

It should be recalled that following the doll choice tasks, subjects were explicitly asked to give the reasons for their choice. Perhaps the most interesting finding with regard to how children describe their doll choices is that they say remarkably little about color. Each of the 192 subjects were asked seven questions about the dolls. Out of a total of 1,344 possible responses, only

TABLE 3

AVERAGE RACIAL IDENTIFICATION SCORES

Sex/race of subject	Age	
	Nursery	Kindergarten
Male		
White	2.67	2.53
Black	2.00	2.71
Female		
White	2.46	2.82
Black	2.67	2.88

66 (5%) were made on the basis of color or other racial characteristics.

It may be that the children were simply inhibited about stating such reasons for choices. If this were so, however, it would be expected that more color descriptions would emerge with a same-race examiner. No differences, however, were attributable to race of examiner. Moreover, if inhibition was involved, the older children should give fewer color responses. Actually, they gave slightly more (again, a nonsignificant difference). More typical of the verbal responses given by children were as follows: His hair is too short; She bothers him too much; He shouts too much, etc., which referred to either gender cues, clothing, or imagined personality characteristics. Thus, it may be concluded that children did not respond overtly to racial cues on the doll task with great frequency.

Discussion

The pattern of results obtained indicates that responses based upon racial cues are already quite complex, even in three-year-old children. Although preschool children are not ordinarily considered to be sophisticated test takers, it was clear that the children were responding to many subtle nuances of the situation. This was particularly clear in the white kindergarten sample, where doll selections apparently were geared to the race of the tester. This ambivalence about expressing attitudes was not characteristic of the older black children who exhibited a slight preference for black dolls.

The overall findings obtained with regard to children's doll choices were clearly not in accordance with the clear-cut preferences for white dolls previously reported by most earlier investigators. It would indeed be tempting to interpret the present findings as indicative of a general lowering of children's prejudicial attitudes. A few recent investigators, also finding that young black children do not show strong white preferences have, in fact, attributed this to both general societal change and the increasing pride associated with black identity (Datcher, Savage, & Checkosky, 1973; Hraba & Grant, 1971). A complete look at research in the area, however, suggests that a more complex explanation may be needed. The Clarks' early findings with dolls have been replicated quite recently in their original community (Asher & Allen, 1969). Moreover, preferences for white characters in picture choice tasks have been found by many investigators (e.g., Williams & Roberson, 1967). It is, of course, possible that black is more beautiful to children in certain parts of the country (e.g., large urban centers) than in others, although no such pattern is easily discernible on the basis of available evidence.

It is the view of the present investigators that the discrepant findings of the present study are more likely due to procedural differences than to either historical change or geographical variations. A number of explanations are possible in this regard. One possibility is that when given another basis for choice (i.e., gender), skin color is not as salient for young children. Since only skin color was varied in the present study, a second possibility is that earlier results reflected a preference for cues other than skin color. Children, like the proverbial "gentlemen," may prefer blondes. A third possibility is that in earlier studies subjects were not actually expressing their own racial preference, but were rather giving what they regarded as the socially desirable response anticipated by the examiner. Thus, responses to dolls of varying skin color may not necessarily be related to behavior outside of the immediate testing situation. To date, no evidence exists to suggest they are. In summary, there appears to be ample reason to question both the validity of some of these earlier findings and whether the task itself is an appropriate indicator of a child's attitude.

REFERENCES

Asher, S. R., & Allen, V. L. Racial preference and social comparison processes. *Journal of Social Issues*, 1969, **25**, 157–165.

Bucky, S. F., & Banta, T. J. Racial factors in test performance. *Developmental Psychology*, 1972, **6**, 7–13.

Clark, K. B., & Clark, M. P. Racial identification and preference in Negro children. In E. E. Maccoby, T. M. Newcomb, & E. L. Hartley

(Eds.), *Reading in social psychology*. New York: Holt, 1947.

Datcher, E., Savage, J. E., & Checkosky, S. F. Investigation of school type, grade, sex and race of examiner on the racial preference and awareness of black and white children. S.R.C.D. paper, Philadelphia, April, 1973.

Greenwald, W. J., & Oppenheim, D. B. Reported magnitude of self-mis-identification among Negro children—artifact? *Journal of Personality and Social Psychology*, 1968, **8**, 49–52.

Hraba, J., & Grant, G. Black is beautiful: A reexamination of racial preference and identification. *Journal of Personality and Social Psychology*, 1971, **16**, 398–402.

Katz, P. A. Stimulus predifferentiation and modification of children's racial attitudes. *Child Development*, 1973, **44**, 232–237.

Katz, P. A., Johnson, J., & Parker, D. Racial attitudes and perception in black and white urban school children. *Proceedings of the 78th Annual Convention of the American Psychological Association*, 1970, **5**, 311–312. (Summary)

Porter, J. *Black child, white child: The development of racial attitudes*. Cambridge, Mass.: Harvard University Press, 1971.

Sattler, J. M. Racial "experimenter effects" in experimentation, testing, interviewing, and psychotherapy. *Psychological Bulletin*, 1970, **73**, 137–160.

Williams, J. E., & Roberson, J. K. A method for assessing racial atttiudes in preschool children. *Educational and Psychological Measurement*, 1967, **27**, 671–689.

(Received January 29, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 669-679

INDIRECT REVIEW AND PRIMING THROUGH QUESTIONS[1]

ERNST Z. ROTHKOPF[2] AND MARJORIE J. BILLINGTON

Bell Laboratories, Murray Hill, New Jersey

Three experiments were performed to determine (*a*) whether attempts to answer a question facilitate the retention of other topically related material that is not directly relevant to the original question (indirect review); and (*b*) whether performance on a test item is better if a question topically related to the test item has recently been asked (priming). Evidence for indirect review was obtained following study of a 6,000-word text using adjunct questions. No priming effects were observed. The indirect review phenomenon suggests that searching one's memory to answer a question strengthens or makes more available a system of semantically related memory features broader than the memory requirements for the initial question.

The purpose of these experiments was to explore two questions: (*a*) Do attempts to recall recently read material aid the retention of other material that is topically related to the initial recall attempts (indirect review); and (*b*) Is performance on a test item better if a question topically related to the test item has recently been asked (priming)?

These experimental problems were prompted by relatively broad issues in conceptualizing those properties of human memory that bear on instruction. Memories evoked by external events such as questions are thought of as broadly reconstructive of the experiences that have produced these memories. Evocations of internal representations are assumed to propagate along dimensions of semantic similarity such as common topics and contexts.

Consequently, it is plausible to argue that external events such as questions can result not only in (*a*) the activation of memory structures specifically required to answer each question but also in (*b*) the activation of internal representations which are topically related to adjunct questions and substantially broader in scope. The activation of other topically related memories implies that the chain of hypothesized internal events resulting from a question can facilitate other subsequent test performances. Facilitation occurs because (*a*) the internal representations, through the principle of exercise, have general instructive properties that produce relatively permanent changes in memory (review); or (*b*) the question temporarily activates topically related domains in memory and therefore makes sufficient responses to other topically related questions temporarily more accessible or more likely (priming).

The direct instructive effect of adjunct questions which has been repeatedly observed (Rothkopf, 1966; Rothkopf & Bisbicos, 1967; Rothkopf & Bloom, 1970) can be understood in terms of the relatively permanent strengthening, through review, of evoked memorial representations directly related to the question. Postreading performance on the questions that have been seen during the course of study was found to be substantially increased even though there was no opportunity for rereading related portions of the text. The direct instructive effect has been found to be greater when informative feedback is provided after answering the adjunct question, but it is of substantial magnitude even when no direct feedback is available to the subject (Rothkopf, 1971). It has been studied mainly in situations where the content of the memory-

[1] We are indebted to Nancy Snidman of Fairleigh Dickinson University in Madison, New Jersey for her help in conducting this experiment.

[2] Requests for reprints should be sent to Ernst Z. Rothkopf, Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974.

evoking question has been identical to the test question on which the effect of exercise was measured (e.g., Rothkopf & Bisbicos, 1967; Rothkopf & Bloom, 1970), where it has been related to the criterion question by a simple syntactic or semantic transformation (Rothkopf & Coke, 1966), or in which the required criterion test information was a mediating term in the logical structure relating the adjunct question to the source text (Frase, 1969).

Evidence for priming has been provided by a recent study by Meyer (1973). He has reported that the time required to judge that a string of letters was an English word was shorter when the presentation of the word was preceded by the judgment of another semantically related rather than an unrelated word. Similar effects involving the likelihood of certain responses have been reported in word association studies (e.g., Martin, 1964).

Some recent observations suggest that it may be possible to measure review or priming effects which were attributable to topical propagation of memory processes evoked by questions beyond the representations needed to answer the question. McGaw and Grotelueschen (1972) reported that post-reading test performance was elevated on questions that have substantial elements in common with adjunct questions but could not be directly answered from the material in the adjunct question. For example, post-reading performance on

Question 1,

The surveying ship, which recovered starfish from a depth of 1,260 fathoms in 1860, was exploring a route for a cable from Faroe to ?.

was increased by seeing the following adjunct question during the course of study:

Question 2,

The surveying ship ?, which recovered starfish from the depth of 1,260 fathoms in 1860, was exploring a route for a cable from Faroe.

Questions 1 and 2 were derived from a text segment that read as follows:

Then from the surveying ship Bulldog, examining a proposed northern route for a cable from Faroe to Labrador in 1860, came another report. The Bulldog's sounding line, which at one place had been allowed to lie for some time on the bottom at a depth of 1,260 fathoms, came up with 13 starfish clinging to it.

McGaw and Grotelueschen's adjunct questions may have provided the occasion for the review or evocation of memory structures which were topically related to the adjunct questions but which were sufficiently broad in scope to facilitate performance on other questions about that topic. It is as if the subject, in searching his memory for an answer to Question 2, also strengthens or makes more available certain memory features that were critical for Question 1 but not for Question 2.

The conditions induced by questions, which dispose toward indirect review closely resemble those required for priming effects. Both the indirect review and the priming hypotheses are invoked to account for the improved performance on a question that is a consequence of an earlier experience with a question on a closely related topic. The critical difference between the two hypothesized processes is the dependence of the priming mechanism on relatively short temporal intervals between the two related questions. Priming effects such as those reported in word association studies (e.g., Martin, 1964) disappear quickly in time. No time effect is assumed for the indirect review effect since the process presumably affects relatively permanent representations in memory.

The first and second of the experiments were aimed at exploring review effects produced by adjunct questions. The second and the third experiment involved attempts to investigate priming effects produced by questions.

Experiment 1

Method

General Procedure

Subjects read a 24-page typewritten passage with two adjunct questions embedded after every 4 pages. The adjunct questions were derived from the material in the 4 pages that had just preceded the questions. Immediately after reading, the subjects were tested. This criterion test included questions closely matched to the adjunct questions as well as some test items that were unrelated to

TABLE 1

DESIGN OF EXPERIMENT 1

Group	AQ	CT ($M + \bar{M}_1$)	AQRT ($AQ + \bar{M}_2$)	Test order
A_1 sequential	A_1	A_2B_2	A_1B_1	sequential
A_1 random	A_1	A_2B_2	A_1B_1	random
B_1 sequential	B_1	A_2B_2	A_1B_1	sequential
B_1 random	B_1	A_2B_2	A_1B_1	random
A_2 sequential	A_2	A_1B_1	A_2B_2	sequential
A_2 random	A_2	A_1B_1	A_2B_2	random
B_2 sequential	B_2	A_1B_1	A_2B_2	sequential
B_2 random	B_2	A_1B_1	A_2B_2	random

Note. The *n* for each group is 15. Abbreviations: AQ = adjunct question, CT = criterion test, AQRT = adjunct question retest, M = items matched to AQs, $\bar{M}_1$ = unmatched items, and $\bar{M}_2$ = items matched to 12 items in the criterion test.

the adjunct questions. A second test covering the adjunct questions was administered subsequently (adjunct question retest).

Materials

An approximately 6,000-word passage from Rachel Carson's *The Sea Around Us* was used.[3] It was typed on 24 full pages. Two questions were constructed for each page for a total of 48 questions. The 2 questions from each page were matched in that they asked for information about closely related elements from that page, frequently from the same sentence, and generally had several phrases in common. However, each pair was constructed so that one question of the pair could not be answered by inspecting the other. An example of a pair of matched questions was given in the introduction.

The matched pairs of questions from each consecutive set of four pages were arbitrarily divided into two groups of pairs, A and B, for a total of 12 pairs in each group. The only restriction on the division of the matched pairs into two groups was that the first, second, third, and fourth pages of the four-page interquestion text segments were equally represented in A and B.

Four sets of 12 questions each were assembled in the following manner: A_1, composed of one member of each matched pair of Set A; A_2, composed of the other member of each pair in Set A; B_1, composed of one member of each matched pair in Set B; and B_2, composed of the other question in each matched pair in Set B.

[3] Permission for the experimental use of these copyrighted materials was kindly granted by the publishers, Oxford University Press, 200 Madison Avenue, New York, New York 10016.

Design

The experimental design is shown in Table 1. One fourth of the subjects were given A_1, A_2, B_1, or B_2, respectively, as experimental adjunct questions. For the subject assigned to A_1 or B_1, the criterion test consisted of the 12 A_2 and the 12 B_2 questions. It should be noted that half of the items of the criterion test were matched to adjunct questions while the remainder were not. Matched and unmatched items were mixed together in a nonsystematic manner. They will be referred to as M and $\bar{M}_1$, respectively. Subsequent to the criterion test, subjects were given another test, adjunct question retest, that consisted of the A_1 and B_1 questions, that is, half of the adjunct question retest was composed of the previously seen adjunct questions, while the remaining items consisted of questions never seen before but which happened to be *matched* to 12 items in the criterion test. The latter will be referred to as $\bar{M}_2$. For subjects assigned A_2 or B_2 as adjunct questions, the criterion test consisted of the questions from Sets A_1 and B_1. These subjects were given the questions from Sets A_2 and B_2 as the adjunct question retest.

For exploratory purposes, the test questions of the criterion test and the adjunct question retest, respectively, were presented to half of the subjects in each adjunct question group in the same ordinal sequence in which the materials underlying these items occurred in the text. For the remaining subjects, the sequential order of items of the criterion test and the adjunct question retest were randomly arranged. Fifteen subjects were assigned to each of the eight adjunct question by test-sequence combinations for a total of 120 subjects.

Procedure

The subjects received a package of four envelopes (see Rothkopf & Coke, 1968, for details of this procedure). The outside of each envelope was marked with a large numeral which uniquely identified it. Envelope Number 1 contained the passage with the embedded adjunct questions; Number 2 contained a short answer test (criterion test); Number 3 contained another short-answer test (adjunct question retest); and Number 4 contained a background questionnaire to determine the subject's familiarity with the materials. Inside each envelope was a distinctively colored sheet giving detailed directions on how to use the materials contained in the envelope. The directions were to study the text carefully and not to reinspect any page that they had already read. Responses to adjunct questions were written into the designated answer space on the question blank and no feedback as to the appropriateness of the answer was provided. A monitor took care to assure that all subjects worked through the envelopes in the proper order.

Study and test time were controlled by the subjects. The subjects were run in small groups of

TABLE 2

AVERAGE NUMBER CORRECT RESPONSES FOR VARIOUS EXPERIMENTAL CONDITIONS WITH SEQUENTIAL AND RANDOM TEST ORDERS ON THE CRITERION TEST AND THE ADJUNCT QUESTION RETEST

Test order	Criterion test		Adjunct question retest	
	Matched (M)	Not matched ($\bar{M}_1$)	Seen (adjunct questions)	Not seen ($\bar{M}_2$)
Sequential	3.88	3.43	5.52	3.40
Random	3.27	2.92	5.87	3.08
M	3.58	3.17	5.70	3.24

4–10 persons. Each subject was seated in a small cubicle.

Subjects

Paid volunteer college students (n = 120) served as subjects. None had 15 or more credit hours of college-level biology (including the current semester), and none stated that they had read *The Sea Around Us* in the recent past.

Results

Criterion Test

The correct responses on the criterion test were subjected to a 4 × 2 × 2 analysis of variance. The factors were the four sets of adjunct question conditions, test sequence (sequential vs. random), and matched versus unmatched items, with repeated measures on the last factor. The data pertinent to this analysis as well as the other main data from the criterion test and adjunct question retest are summarized in Table 2. The average number of correct responses on matched (M) items ($\bar{X}$ = 3.58, σ = 2.01) over all conditions was significantly greater (F = 7.26, df = 1/112, $p < .01$) than that for unmatched test questions, $\bar{M}_1$ ($\bar{X}$ = 3.17, σ = 1.77). This result was consistent with the findings reported by McGaw and Grotelueschen (1972).

Performance on test questions sequenced in the same order as the text ($\bar{X}$ = 3.66, σ = 1.93) was higher than on random arrangements of test items ($\bar{X}$ = 3.09, σ = 1.83), but the difference between the two conditions was not significant (F = 3.41, df = 1/112, $.10 > p > .05$). The somewhat more accurate recall for the nonrandom test order suggests the possibility of some priming effects since questions from contiguous text segments tend to be more closely related topically than randomly ordered questions. There were no significant differences among the four adjunct question sets nor were there any significant interactions.

Adjunct Question Retest

The adjunct question retest consisted of 24 test questions. Of these, 12 had been seen before during the course of study as adjunct questions. A 4 × 2 analysis of variance was performed on correct responses on the adjunct question retest. The factors were the four adjunct question sets and whether the items had been seen or not as adjunct questions, with repeated measures on subjects on the last factor. Previously seen adjunct question retest items (adjunct questions) produced more correct responses ($\bar{X}$ = 5.70, σ = 2.65) than those not previously seen, $\bar{M}_2$ ($\bar{X}$ = 3.24, σ = 1.89; F = 96.42, df = 1/116, $p < .01$). The adjunct question set factor and the interaction were not significant.

Interval Effects

There was some suggestion in the present data that the time interval between inspection of the relevant text segment and adjunct question has some bearing on the indirect review effect. The data relevant to this conjecture are given in Table 3, which was derived from the following analysis. The test questions used in the criterion test were divided into four groups of 12 each depending on whether they were from the first, second, third, or fourth pages of the four-page text segments that preceded adjunct questions in the text. The average number of correct responses per item was calculated for each page location. This was done separately for the M and the $\bar{M}_1$ items. The results, tabulated respectively in Columns 2 and 3 of Table 3, indicate that the indirect review effects produced by a related (matched) adjunct question were strongest when adjunct questions were relevant to the text page that immediately precedes them.

TABLE 3

AVERAGE CORRECT RESPONSES ON THE CRITERION TEST FOR ITEMS FROM FOUR LOCATIONS IN EACH FOUR-PAGE TEXT SEGMENT

Page	Matched (M)[a]	Not matched ($\bar{M}_1$)[a]	Difference
1	9.25	10.00	−.75
2	7.67	6.33	1.34
3	8.00	6.67	1.33
4	10.83	8.75	2.08

[a] The figures shown in these columns are averaged over test items rather than over individual subjects. The column means, which would be based on 48 items, may be converted by multiplying with 48/120 to the matched and unmatched item data described in the first paragraph of the Results section. The latter have been averaged over 120 subjects.

A 2 × 2 analysis of variance, comparing data from the most distal (p. 1) and the most proximal (p. 4) pages, under the matched and unmatched conditions, support this conclusion. Item means rather than subjects were used for the purposes of this analysis. For the matching treatment, the obtained F was not significant ($F = .97$, $df = 1/22$), but the interaction between matching and page location was ($F = 4.39$, $df = 1/22$, $p < .05$). A t test between matched and unmatched items from Page 4 locations was significant ($t = 2.18$, $df = 22$, $p < .05$). The matched and unmatched condition did not differ significantly for Page 1 locations.

Discussion

Experiment 1 confirmed McGaw and Groteluesch en's (1972) observation that adjunct questions in the postreading position can produce an indirect review effect. The replication was obtained with materials that were substantially identical to those used by McGaw and Groteluesch en, but it involved a larger number of test items and experimental procedures which differed in several ways from those used in their study. The review effect became manifest through facilitated performance on test items that were topically related to the adjunct questions. The effect was small. The performance facilitation was 12.8% of the comparison score and an increase of only 4% in the absolute proportion of correct responses.

EXPERIMENT 2

The experimental passage and questions used in Experiment 1 were printed in mimeographed booklets. Even though the subjects were closely monitored, the possibility could not be absolutely ruled out that the indirect review effect may have been produced by subjects surreptitiously leafing back to relevant portions of the text while trying to answer the experimental questions or immediately thereafter. In doing so, they may also have inspected materials relevant to the test item that was matched to the adjunct question. The finding that the review effect was larger when the relevant materials are within one or two pages of the experimental question (see Table 3) would be consistent with this conjecture. Alternatively, this result may be an indication that the review effect operates only on representations in primary memory (see Waugh & Norman, 1965) and that these decay or become less accessible in time.

Experiment 2 was therefore performed in order to test for the indirect review effect when surreptitious reinspection of the text following the adjunct question was precluded. For this purpose, all experimental material was presented on photographic slides under conditions that prevented the reexamination of the text after the initial inspection. A second purpose of Experiment 2 was to explore priming effects associated with the successive presentation of two related questions.

Method

Materials and General Procedure

A negative 35-millimeter slide was prepared of each of the 24 typed pages of the 6,000-word experimental passage used in Experiment 1. The slides were projected one at a time on a screen. Inspection time was controlled by the subject, but the subject could not return to any previously inspected slide. One adjunct question was included in the slide sequence after every second text page. Immediately after reading, a 24-item test was administered consisting of a sample of (*a*) questions matched to adjunct questions; (*b*) questions not matched to any adjunct questions; (*c*) pairs of matched questions, one of which had been used

as an adjunct question; and (*d*) pairs of matched questions neither of which had been used as an adjunct question.

Adjunct Questions and Test

The two sets of 12 matched questions (A_1:A_2; B_1:B_2) employed in Experiment 1 were used. One fourth of the subjects used A_1, A_2, B_1, and B_2, respectively, as adjunct questions (AQ). One adjunct question was administered every two text slides (pages). It was derived from the two pages which preceded its administration.

Immediately after reading, a 24-item short-answer test was administered. This consisted of (*a*) four questions (M) matched to adjunct questions used during reading but without the matched adjunct questions occurring in the test; (*b*) four questions ($\bar{M}$) for which no matching adjunct questions had been administered during reading and for which no matching questions were presented in the test; (*c*) four *pairs* of matched questions, one member of each pair had been used as an adjunct question (AQ_P) which was followed immediately in the test sequence by the question that was matched to it (M_P); and (*d*) four *pairs* of matched questions neither of which had been used as adjunct questions; the first member of each matched pair in the test sequence was designated as $\bar{Q}_P$ and was immediately followed by the question that was matched to it ($\bar{M}_P$). The questions were presented in the same order on the test as the sequence of the source material in the text. As a consequence, the matched pair of questions, AQ_P:M_P or $\bar{Q}_P$:$\bar{M}_P$ were always presented in immediate sequence. The review effect may be inferred if questions from Category M resulted in more correct responses than those from Category $\bar{M}$. Priming may be inferred if $\bar{M}_P > \bar{M}$ and if $M_P > M$ with respect to correct responses. It may be concluded that adjunct questions had direct instructive effects if AQ_P resulted in more correct responses than $\bar{Q}_P$.

The equal use of the four question groups A_1, A_2, B_1, B_2, as adjunct questions and the use of three different test arrangements for each adjunct set made it possible to have all 48 available questions appear equally often in the various classifications of test items.

Apparatus

Each subject was seated in a booth that was fitted for rear projection of the text material. The projector advance was operated by the subject with a hand switch. The circuits connecting the subject's switch to the projection equipment were wired through control panels in the experimenter's booth. This allowed the experimenter to monitor each subject's progress through the text. For further details of the apparatus see Rothkopf and Bloom (1970).

Procedure

The subject read the passage which was projected, one page at a time, on the screen in front of him. After every two text slides there was a question slide distinguished by a green background field. The subject wrote his answer to the question on the answer sheet provided. The experimenter gave the subject the written, 24-item criterion test after the subject had read the entire passage.

Study and test time were controlled by the subject. The subjects were run in groups of from two to three persons.

Subjects

Paid volunteer students (n = 180) served as subjects. Of these, 144 were college students and 36 were high school seniors. None had read *The Sea Around Us* in the recent past, and none had 15 or more credit hours of college biology.

Results

Review Effect

Over all 180 subjects, each of the 48 possible test questions was assigned 15 times to each of the several experimental roles. The analytical comparisons of this experiment were based on correct responses per test question averaged over all 48 items. The review hypothesis was tested by comparing performance between M and $\bar{M}$ items. The results were consistent with the review hypothesis ($\bar{X}_M = 6.08 > \bar{X}_{\bar{M}} = 5.44$; maximum score = 15). The effect is small but statistically reliable. Reliability was tested by comparing results for each of the 48 possible test items when it was assigned to the M and $\bar{M}$ role in the test. Performance on 28 cases was greater for M than for the corresponding item in the $\bar{M}$ state; $\bar{M}$ was greater than M in 14 instances; and the two conditions were exactly equal in 6 cases. A sign test indicates that results such as these would occur by chance with $p < .046$ ($z = 2.01$). These results therefore confirm those of Experiment 1. Performance on a test item was facilitated if a question that is topically related had been asked during study of the text.

Priming

Two comparisons relevant to the priming hypothesis were possible in this experiment.

These were (*a*) when the priming item in testing had also been used as an adjunct question during reading and required the comparison of M_P with M and (*b*) when the priming item was *not* used as an adjunct question, that is, the comparison between $\bar{M}_P$ and $\bar{M}$. The latter comparison could also be made by contrasting the performance between $\bar{Q}_P$ and $\bar{M}_P$. The $\bar{Q}_P$ items always preceded $\bar{M}_P$ in the test sequence and for that reason $\bar{Q}_P$ may be thought as unprimed while $\bar{M}_P$ is primed. However, since the test was presented in booklet form, the order in which the items were responded to was not entirely certain and the nature of the $\bar{Q}_P$:$\bar{M}_P$ contrast was therefore less clear than $\bar{M}$:$\bar{M}_P$.

As indicated previously, the average number of correct responses for M items was 6.08 while the mean number of correct responses for M_P was 6.10. It may therefore be concluded that there is little evidence for priming in items for which an adjunct question was presented during inspection of the text.

On the other hand, the data indicate the possibility of priming effects in test items without matching adjunct questions during the study of the text. The average number of correct responses for $\bar{M}$ items was 5.44 and for $\bar{M}_P$ it was 6.08. Procedures, similar to the method used in the review comparison, involving individual test items were again used. The results were as follows. For 27 test items, $\bar{M}_P > \bar{M}$ and in 12 cases $\bar{M} > \bar{M}_P$. Performance was exactly equal on $\bar{M}$ and $\bar{M}_P$ items in 9 cases. A sign test indicates $p < .05$ ($z = 2.24$). This result must be interpreted with some caution since the $\bar{M}$ data have been used in a previous comparison. Some further corroboration for the conclusion that priming enhances performance in items not matched by any adjunct question may be found by noting that the average number of correct responses to $\bar{Q}_P$ items ($\bar{X} = 5.69$) was also somewhat lower than for $\bar{M}_P$ ($\bar{X} = 6.08$). However, in only 21 cases $\bar{M}_P > \bar{Q}_P$, in 17 cases $\bar{Q}_P > \bar{M}_P$, and for 10 items $\bar{M}_P = \bar{Q}_P$ and the difference was not significant.

Finally, it is worth noting that, as in Experiment 1, substantial direct instructional effects of questions were observed. The powerful direct instructional effect of questions was indicated by the high performance on AQ_P items ($\bar{X} = 8.13$). When contrasted with the $\bar{Q}_P$ question ($\bar{X} = 5.69$), $AQ_P > \bar{Q}_P$ for 33 items, $\bar{Q}_P > AQ_P$ in 11 cases, and $AQ_P = \bar{Q}_P$ for 4 items ($z = 3.17$, $p < .002$).

Discussion

The results of Experiment 2 indicate that a response to a question improves performance on other topically related items. The finding that performance on M was higher than that on $\bar{M}$ replicates the review effect observed in Experiment 1 under conditions that rule out the possibility of surreptitious reexamination of the text. However, in view of the small magnitude of this effect, the phenomenon is probably of more theoretical than practical importance.

There were no observable test sequence effects of performance when an item was preceded by a matching question that had been used as an adjunct question (i.e., M_P = M). There was some indication however that test sequence facilitation may have taken place for those items that were preceded in the testing sequence by matched questions that were *not* used as adjunct questions (e.g., $\bar{M}_P > \bar{M}$ or $\bar{M}_P > \bar{Q}_P$). A priming account can be accommodated to these results. It would be more parsimonious, however, to account for these data in terms of indirect review. Such indirect review is assumed to take place during attempts to answer $\bar{Q}_P$ and it improves the subject's ability to respond to $\bar{M}_P$, the item matched to $\bar{Q}_P$ which follows immediately after it in the testing sequence.

Experiment 3

Experiments 1 and 2 have provided reasonably firm evidence that an attempt to answer a question can improve the subsequent ability to answer other questions from the same topico-spatial neighborhood of the text. Whether this is interpreted as an indirect review or priming phenomenon de-

pends on the effect of the time interval between the two related questions. If the facilitating effect lasts for a substantial period of time, that is, if the first questions have produced a relatively permanent change in memory, then an indirect review interpretation can be invoked. Experiments 1 and 2 provided substantial indications of indirect review phenomena. The priming hypothesis depends on the demonstration of a weakening of the facilitation effect as the temporal interval between the two related (matched) questions is increased. The data of Experiment 2 were equivocal with respect to the priming hypothesis, but priming continues to be an interesting conjecture about facilitation between two related questions because priming effects have been obtained in memory search (Meyer, 1973) and in word association experiments (Martin, 1964).

In Experiment 2, test items preceded in the test sequence by a matched question that had not been an adjunct question resulted in somewhat higher performance than the questions that preceded them and also resulted in higher performance than unmatched, unprimed items. This suggests that priming may have played some role in supporting subjects' test performance, but it also leaves open the possibility that the higher performance on the second question may have resulted from indirect review that took place during testing.

Experiment 3 was intended to explore the priming factor by systematically manipulating the interval between priming events and subsequent related items in testing.

Method

Materials and Procedure

Materials and procedure were exactly as in Experiment 2. Subjects read a 24-page passage which was projected one page at a time onto a screen in front of them. An experimental question (adjunct question) was embedded after every second page. The adjunct question was derived from information in the two pages immediately preceding it. Immediately after reading the passage, the subjects were tested. This criterion test (Criterion Test 1) included some adjunct questions as well as questions that were closely matched to the adjunct questions. A second criterion test (Criterion Test 2) similar to the first was administered either 5 or 30 minutes after the subject completed the first one. It included items matched to questions in Criterion Test 1. During this delay period, the subjects judged comprehension difficulty of a number of short, unrelated paragraphs.

Tests

Criterion Test 1 included 24 items. It consisted of (*a*) four *pairs* of matched questions; one member of each pair had been used as an adjunct question, (it will be referred to as AQ_{IP}; adjunct questions, immediate priming) which was followed immediately in the test sequence by the question that was matched to it (M_{IP}); (*b*) four *pairs* of matched questions neither of which had been used as adjunct questions; the first member of each matched pair in the test sequence was designated $\bar{Q}_{IP}$ and was immedately followed by the question that was matched to it ($\bar{M}_{IP}$); (*c*) four questions that had been used as adjunct questions (referred to as AQ_{DP}; adjunct questions, delayed priming), the questions matched to these were presented in Criterion Test 2; (*d*) four questions ($\bar{Q}_{DP}$) which were not used as adjunct questions, with their matching questions presented in Criterion Test 2.

Criterion Test 2 consisted of 24 items not used in Criterion Test 1. These comprised (*a*) four pairs of matched questions, one member of each pair had been used as an adjunct question (it will be referred to as $AQ_{IP\text{-}2}$; adjunct questions, immediate priming, Test 2) which was followed immediately in the sequence by the question that was matched to it ($M_{IP\text{-}2}$); (*b*) four *pairs* of matched questions neither of which had been used as an adjunct question; the first member of each matched pair in the test sequence was designated $\bar{Q}_{IP\text{-}2}$ and was immediately followed by the question that was matched to it ($\bar{M}_{IP\text{-}2}$); (*c*) four questions ($M_{DP\text{-}2}$) each matched to one of the AQ_{DP} in Criterion Test 1; and (*d*) four questions ($\bar{M}_{DP\text{-}2}$) each matched to a $\bar{Q}_{DP}$ item in Criterion Test 1.

Performance on M_{IP} and $M_{IP\text{-}2}$ allows assessment of immediate priming by questions seen during reading (adjunct question); performance on $\bar{M}_{IP}$ and $\bar{M}_{IP\text{-}2}$ should reflect immediate priming by related questions ($\bar{Q}$) not seen during reading. Responses to $M_{DP\text{-}2}$ and $\bar{M}_{DP\text{-}2}$ make it possible to examine any effects due to delay in priming. These two classes of items were removed from the matched adjunct questions or $\bar{Q}$ questions by the interval that separates Criterion Test 1 and Criterion Test 2, that is, either 5 minutes or 30 minutes.

One fourth of all subjects were assigned question groups A_1, A_2, B_1, B_2, respectively, as adjunct questions. Four test arrangements were used for each adjunct question assignment. This made it possible to equalize the frequency with which each criterion test question was used in each of the four categories.

Subjects

Paid volunteer high school students ($n = 72$) served as subjects. None had read *The Sea Around*

TABLE 4

MEAN CORRECT RESPONSES ON TEST ITEMS OF VARIOUS TYPES FOR THE TWO TREATMENTS IN EXPERIMENT 3

Test item	Test 1			Test item	Test 2		
	Treatment				Treatment		
	5 minutes	30 minutes	$\bar{X}$		5 minutes	30 minutes	$\bar{X}$
AQ_{IP}	1.92	1.94	1.93	AQ_{IP-2}	1.61	2.08	1.85
M_{IP}	1.42	1.67	1.54	M_{IP-2}	1.39	1.56	1.47
$\bar{Q}_{IP}$	1.50	1.39	1.44	Q_{IP-2}	1.39	1.61	1.50
$\bar{M}_{IP}$	1.28	1.56	1.42	$\bar{M}_{IP-2}$	1.11	1.78	1.44
AQ_{DP}	1.75	2.20	1.97	M_{DP-2}	1.33	1.53	1.43
$\bar{Q}_{DP}$	1.14	1.56	1.35	$\bar{M}_{DP-2}$	1.00	1.50	1.25

Note. Abbreviations: AQ_{IP}, AQ_{IP-2} = adjunct questions which served as immediate primers; M_{IP}, M_{IP-2} = matched items which were immediately preceded by AQ primers; $\bar{Q}_{IP}$, $\bar{Q}_{IP-2}$ previously unseen questions which served as immediate primers; $\bar{M}_{IP}$, $\bar{M}_{IP-2}$, items which were immediately preceded by $\bar{M}$ primers; AQ_{DP}, $\bar{Q}_{DP}$ adjunct questions and previously unseen questions, respectively, which served as delayed primers; M_{DP-2} and $\bar{M}_{DP-2}$ items matched to AQ_{DP} and $\bar{Q}_{DP}$ primers.

Us in the recent past. Half of these were assigned in a nonsystematic manner to the 5-minute and the 30-minute delay condition.

Results

The results from the two criterion tests are summarized in Table 4. They must be interpreted cautiously since each of the treatment cells represents only 144 observations.

The major comparisons of interest were immediately primed (M_{IP-2}, $\bar{M}_{IP-2}$) and delayed primed items (M_{DP-2}, $\bar{M}_{DP-2}$) in the second criterion test. In the four relevant comparisons that were possible over the two treatments, the primed items produced slightly higher performance than the delayed primed items in three out of four cases. For purposes of statistical comparison, scores on M_{IP-2} were combined with $\bar{M}_{IP-2}$ as were scores on M_{DP-2} with $\bar{M}_{DP-2}$. A 2 × 2 analysis of variance (Treatment × Priming Delay) with repeated measures on the second factor was performed on the combined scores. The analysis of variance indicates that the 30-minute group produced more correct responses than the 5-minute group ($F = 5.60$, $df = 1/70$, $p < .05$). This was probably due to chance sampling differences since, as may be noted from Table 4, the 30-minute treatment outperformed the 5-minute group for five out of six test item types on the immediate tests (Criterion Test 1) when the two treatments were operating under the same conditions. Performance in Criterion Test 2 was higher for primed items in both treatments but the priming factor was not significant ($F = 1.14$, $df = 1/70$) nor was the interaction ($F = .10$, $df = 1/70$).

Inspection of these data also indicated that little or no retention loss took place during either the 5-minute or 30-minute delay interval.

Another test of the priming hypothesis is possible using the results from Criterion Test 1. According to the priming hypothesis, correct responses to $\bar{M}_{IP}$ items should be more numerous than to $\bar{Q}_{DP}$ questions. For the 5-minute group, this is the case but the difference is not statistically reliable. For the 30-minute group, the average number of correct responses to $\bar{M}_{IP}$ and $\bar{Q}_{DP}$ questions is equal.

Performance on M_{IP} would be superior to $\bar{Q}_{DP}$ if either the review or the priming factor or both were operative. The means for M_{IP} under both the 5-minute and 30-minute treatment were higher than for $\bar{Q}_{DP}$. Combining the two treatments and using the item-by-item comparison and sign test used in Experiment 2, M_{IP} was greater than $\bar{Q}_{DP}$ in 19 cases and smaller in 10 ($z = 1.49$, $p < .07$, one-tailed). It must therefore be concluded that neither review nor priming factor could clearly be demonstrated in this case.

General Discussion

The results of the experiments reported here support the hypothesis that being questioned about a narrow topic while reading enhances the recall of other material closely related to that topic. The data are consistent with the conception that the subject searches his memory while trying to answer an adjunct question during reading. In doing so, the subject reviews and strengthens previously established related memory representations or makes them in other ways more accessible in subsequent tests.

The relationship between the performances facilitated by this indirect review process and the nature of the facilitating adjunct question requires further exploration. In the present studies, the matched questions were always drawn from the same topico-spatial[4] neighborhood of the text, but it is at least in principle possible that either topical or spatial factors may be sufficient to propagate the review effect. The absolute magnitude of the indirect review effect, as observed here and also as reported by McGaw and Grotelueschen (1972), was extremely small. This suggests that the matching operation which was done in a crude unsystematic manner needs to be characterized in more explicit terms. This is required in order to understand more deeply the dimensions along which the indirect review effect operates.

There are several weak suggestions that some priming occurs in testing in the present experiments. However, the experimental facts do not offer sufficient support for this conjecture and it must be rejected. The results that are particularly fatal for the priming hypothesis were that the "priming" effects were not substantially attenuated by interposing 30 minutes between matched test items. It therefore seems more reasonable to conclude that indirect review produced the elevation in $\bar{M}_P$ items in Experiment 2 and the nearly identical performance level on M and $\bar{M}$ both under immediate and delayed priming conditions in Experiment 3. The empirical principle may be stated as follows: A question about any given topico-spatial neighborhood of a text tends to facilitate performance on test items drawn from other text elements topically related to that neighborhood.

An apparent exception to this rule is the absence of elevated performance in $\bar{M}_2$ items on the adjunct question retest in Experiment 1. These items were preceded in the earlier test (Criterion Test) by matched questions ($\bar{M}_1$) and should therefore have benefited by indirect review in attempts to answer $\bar{M}_1$. However, any conclusion that facilitation took place involves here a comparison of $\bar{M}_2$ with $\bar{M}_1$. This comparison is confounded by the longer retention interval for $\bar{M}_2$ items and by possible destructive effects of testing. Consequently, failure to find facilitated performance on $\bar{M}_2$ is not critical for the indirect review hypothesis.

The priming hypothesis should not be ruled out on the basis of the present experiments alone. The common topical threads which ran through most of the 6,000-word experimental passage prevent a sensitive test of the priming hypothesis. A more powerful test would involve a passage with many distinctively different topical themes.

[4] Although the spatial factor was not systematically manipulated, it cannot be ruled out as a contributor to the indirect review effect. This is because in the present experiment, topical relationships among text elements were always accompanied by spatial contiguity.

REFERENCES

Frase, L. T. Structural analysis of the knowledge that results from thinking about text. *Journal of Educational Psychology Monograph*, 1969, **60** (6, Pt. 2).

Martin, J. G. Word-association frequency and the proximity effect. *Journal of Verbal Learning and Verbal Behavior*, 1964, **3**, 344–345.

McGaw, B., & Grotelueschen, A. The direction of the effect of questions in prose material. *Journal of Educational Psychology*, 1972, **63**, 580–588.

Meyer, D. E. Correlated operations in searching stored semantic categories. *Journal of Experimental Psychology*, 1973, **99**, 124–133.

Rothkopf, E. Z. Learning from written materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, 1966, **3**, 241–249.

Rothkopf, E. Z. Experiments on mathemagenic behavior and the technology of written instruction. In E. Z. Rothkopf & P. E. Johnson (Eds.), *Verbal learning research and the technology of written instruction*. New York: Columbia University, Teachers College Press, 1971.

Rothkopf, E. Z., & Bisbicos, E. Selective facilitative effects of interspersed questions on learning from written material. *Journal of Educational Psychology*, 1967, **58**, 56–61.

Rothkopf, E. Z., & Bloom, R. D. Effects of interpersonal interaction on the instructional value of adjunct questions in learning from written material. *Journal of Educational Psychology*, 1970, **61**, 417–422.

Rothkopf, E. Z., & Coke, E. U. Variations in phrasing, repetition intervals, and the recall of sentence material. *Journal of Verbal Learning and Verbal Behavior*, 1966, **5**, 86–89.

Rothkopf, E. Z., & Coke, E. U. Learning about added sentence fragments following repeated inspection of written discourse. *Journal of Experimental Psychology*, 1968, **78**, 191–199.

Waugh, N. C., & Norman, D. A. Primary memory. *Psychological Review*, 1965, **72**, 89–104.

(Received November 30, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 680–687

SEX DIFFERENCES IN COMPREHENSION OF HIGH- AND LOW-INTEREST READING MATERIAL[1]

STEVEN R. ASHER[2] AND RICHARD A. MARKELL

University of Illinois

Previous research has found that elementary-school-age boys read more poorly than girls. The present study investigated whether sex differences in reading comprehension are affected by variations in the interest level of the material. Fifth-grade children's interests were assessed using a picture-rating technique. Each child then read material that corresponded to his or her high- and low-interest areas. The cloze procedure was used to measure comprehension. Results indicated that boys read as well as girls on high-interest material but that they were significantly poorer readers of low-interest material.

There is considerable evidence that elementary school boys read more poorly than girls. Boys lag as much as one half year behind girls on reading achievement tests (Asher & Gottman, 1973; Stroud & Lindquest, 1942), and between 60% and 90% of elementary school children referred for remedial reading instruction are male (Blom, 1971). Although the poorer performance of boys may reflect lack of skill, it is possible that boys are simply less motivated to engage in reading activity. There is evidence that reading has less appeal for boys. For example, boys view reading as a sex-inappropriate activity (Kagan, 1964; Stein & Smithells, 1969) and have a more negative attitude toward reading (Neale, Gill, & Tismer, 1970).

The purpose of the present study was to evaluate the extent to which variations in motivation affect sex differences in reading comprehension. A powerful motivational variable may be the interest level of the material. One study of children's reading books found that much of the material was thematically appropriate for neither boys nor girls (Blom, Waite, & Zimet, 1968). If boys are particularly motivated by the interest level of material, then their performance could be strongly affected by variations in reading material.

There is little research on the relationship between the interest value of reading material and reading comprehension. Data from available studies are inconsistent, perhaps as a result of problems in research design. In one study (Shnayer, 1967), children's interest in the material was assessed after they had read the material and had been tested for comprehension. Since children may be more interested in material they can comprehend as well as in material with topic appeal, the postreading interest measure is of questionable validity. In other studies (Bernstein, 1955; Klein, 1969; Stanchfield, 1967), children were given reading material based on previous studies of children's interests. Since individual children's interests differ from group norms, this procedure introduces experimental error. Finally, previous studies have used reading achievement tests specifically developed for each study with no prior demonstration of test reliability or validity. In many cases, item selection appeared to be arbitrary.

To overcome these limitations, the present study assessed individual children's interests independently of any reading material. In addition, a reliable and valid measure of

[1] The authors wish to thank Marcia Anderson for her help in various phases of the experiment and the Champaign, Illinois, public schools for their cooperation and support. Carol Dweck and John Gottman made valuable comments on earlier drafts of the manuscript.

[2] Requests for reprints should be sent to Steven R. Asher, Department of Educational Psychology, University of Illinois, Urbana, Illinois 61801.

reading comprehension was employed. The experiment consisted of three phases. In the first phase, a picture-rating technique was used to assess children's interests. Students were shown a series of 25 photographic slides representing a wide range of topics. They rated the interest value of each slide. Based on the ratings, three high- and three low-interest areas were identified for each student.

One week later, in the second phase of the experiment, reading comprehension was assessed. Each child read three passages with topics that corresponded to high-interest areas and three passages that corresponded to low-interest areas. Passages were presented in cloze format (Taylor, 1953); every fifth word was deleted from each paragraph and the child's task was to supply each of the missing words. Previous research with the cloze procedure has found that it is reliable and that it correlates highly with standardized reading comprehension achievement tests (Bormuth, 1967, 1968; Rankin & Culhane, 1969). Furthermore, the clearly specifiable rules for creating a cloze passage eliminate the subjectivity and arbitrariness inherent in many other approaches to item construction.

In the third phase of the experiment, students were asked how much they would like to read more about each paragraph topic. This served to assess the validity of the picture technique for selecting material. If the technique is valid, then children should prefer the paragraphs on topics corresponding to high-interest pictures.

In addition to paragraph comprehension and paragraph preference data, standardized reading achievement test data were gathered from student records. From these data, it was determined whether the present sample of subjects shows a significant sex difference in reading achievement test performance. Also, by correlating achievement test scores with cloze scores, an additional check on the validity of the cloze was provided. Finally, the achievement test scores were used to determine whether the effect of interest is different for high- versus low-achieving students.

Method

Subjects

Eighty-seven fifth-grade children, 49 boys and 38 girls, participated in the study. Fifth graders served as subjects since recent evidence (Asher & Gottman, 1973) indicates that sex differences are likely to be found in reading achievement at this grade level. The children, with the exception of four who were repeatedly absent, constituted the entire fifth-grade population of a Champaign, Illinois, public school. Most were from middle-class homes. The children's average IQ on a school-administered Scholastic Testing Service test was 109.

Materials

Interest slides. Color slides were made by photographing pictures from magazines and books. Pictures were selected to represent a wide range of interests for boys and girls. To ensure that each picture represented a single theme, the slides were shown to 34 fifth-grade children in another school. Children were asked to identify the topic of each slide; a naive judge sorted the topics assigned to each slide. Pictures used in the present study were given similar topics by 75% or more of the children. The 25 picture topics are listed below. This order of presentation was randomly selected.

1. Forest
2. Jet Airplane
3. Priest
4. Dogs
5. Astronaut
6. Bride
7. Calf
8. Basketball Players
9. Butterflies
10. Marionettes
11. Monkey
12. Flowers
13. Bullfighting
14. Skiing
15. Food
16. Living Room
17. Maps
18. Painting
19. Circus
20. Race Cars
21. Canoe
22. Model Trains
23. Mother and Child
24. Insects
25. Cats

Reading material. Twenty-five passages corresponding to the various picture topics were selected from the Britannica Junior Encyclopaedia (1970). This source is written for elementary school children in the fourth grade or older (Walsh, 1973) and provides a wide range of topics

in a more consistent style than would be obtained from diverse sources. The material was selected by a person who was unaware of the purpose of the experiment. He was given the list of picture topics and asked to locate a passage for each of the topics. Ten words were deleted from each passage. A 10-space line replaced each deleted word. The first deletion was the ninth word in the passage; subsequent deletions were every fifth word. One entire sentence appeared after the final deletion.

Since the children did not choose their own reading material, it is unlikely that the objective difficulty of passages would vary by sex or interest condition. To test the hypothesis that passages were of similar difficulty across conditions, readability coefficients (Dale & Chall, 1948) were calculated. The Dale-Chall readability coefficient is a function of the average sentence length of the passage and the number of words in the passage not found on a list of familiar words. The average paragraph readability raw score was 6.53; this corresponds to a seventh-grade equivalent score. High- and low-interest readability coefficients were calculated for each child by summing the raw scores for the three high- and three low-interest passages. A 2 × 2 (Sex × Interest) analysis of variance performed on summed readability raw scores indicated no significant sex ($F = .13$, $df = 1/85$), interest ($F = 3.18$, $df = 1/85$), or sex by interest ($F = 1.30$, $df = 1/85$) effects. Thus, the objective difficulty of paragraphs did not vary across conditions.

Procedure

The interest and comprehension tests were administered in two separate sessions one week apart. The children were tested in classrooms during their reading period. To minimize the possibility that children would perceive the connection between the two activities, different experimenters conducted the two sessions. This procedure was apparently effective; in discussions with children following the experiment, only one child inquired about the relationship between the pictures and the reading material.

Interest Assessment

Experimenter 1 told the children, "I'd like to find out about what kids are interested in. I'm going to show you 25 slides. For each slide I'd like you to mark, on the sheets we'll give you, how interesting the picture is to you. Who knows what 'interesting' means?" After a few children had responded, Experimenter 1 summarized their comments by saying, "So, something is interesting when you like it and would like to find out more about it." Experimenter 1 then distributed to each subject a form with twenty-five 1–7 rating scales, and drew a 1–7 scale on the blackboard. At the low end of each scale were the words "not at all interesting" and, at the high end, "very interesting." The nature and use of the rating scale were explained:

> If a picture is very interesting to you—if you like it very much and want to know more about it—mark a number at this end of the scale. [The experimenter pointed to Numbers 5, 6, and 7 of the scale on the blackboard.] You can mark it with a circle, an X, a check, or whatever you want. If a picture is not at all interesting to you—if you don't like it and wouldn't care to find out more about it—mark a number at the low end of the scale. [The experimenter pointed to the Numbers 1, 2, and 3 of the scale.] If the picture is of medium interest to you—if you like it but don't like it a lot—mark a number here. The experimenter pointed to Numbers 3, 4, and 5.] Let's try an example for practice. If I showed a picture of a pile of dollar bills, what number would you choose? [The experimenter called on several students.] If I showed a picture of a piece of dirt, what number would you choose? [The experimenter again called on several students.] So you can see that different people are interested in different things. If anyone has any questions, raise your hand and I'll try to answer them. [Experimenter 1 then presented the slides announcing the number of each one as it was projected.] Here's Picture Number 1... Here's Picture Number 2 . . . , etc.

The slides were presented at the rate of approximately one every 10 seconds. When all slides had been rated, the subjects were asked to write their names on their rating sheet.

Reading comprehension task. One week after the interest assessment, Experimenter 2 gave the children six passages to read. Three of the passage topics corresponded to a child's three highest interest ratings and three to his or her lowest interest ratings. When slides shared the third highest or third lowest rating, topics were randomly selected from those sharing equal ratings. Each of the six passages, appropriately titled in upper-case letters, was mimeographed on 8½ × 11 inch paper and enclosed in a legal-size envelope. The envelopes were numbered from one to six to specify the order in which passages should be read. Forty-three of the children, randomly selected, read the passages in a high–low–high–low–high–low interest sequence while the other 44 children read them in a low–high–low–high–low–high sequence. The particular positions of the three high- and three low-interest passages within these two arrangements were randomly determined. In addition to these six envelopes, each child received a seventh envelope which contained the six reading preference rating scales.

Before the subjects were given the envelopes, Experimenter 2 gave the following instructions:

> I am going to show you a reading game. [Experimenter 2 gave each child a sample paragraph.] This is a paragraph with some words missing. The idea is to read the paragraph and decide what words are missing. Each paragraph has 10 missing spaces. Take a minute to look at

the paragraph. [The experimenter paused.] OK. Now I'll read the paragraph with all of the words in it. You follow along with me. [The experimenter read the sample paragraph aloud, collected the sample paragraph from each subject and then gave each subject the test envelopes.]

You now have seven envelopes. Six have paragraphs in them. Start with the first paragraph and try to fill in all of the missing words. When you are done with a paragraph, put it back in the envelope and put it aside on your desk. Then you can go on to the second envelope; then the third, fourth, fifth, and sixth. Once you put a paragraph in the envelope you can't go back. Do you have any questions?

OK. Read each paragraph carefully and try to fill in the missing words. I can't help you read any of the words, but if you have trouble spelling any words raise your hand and I will help. Spelling doesn't count in this game. If you are having trouble, don't get stuck. Go on to the next part of the paragraph or a new paragraph. You have 40 minutes for the six paragraphs. That should be plenty of time. Any questions?

When you are done with the six paragraphs, open the seventh envelope. It contains some questions about how much you want to read more about each of the topics. If you would like to read more about it, circle one of the high numbers. If you wouldn't like to read more about it, circle one of the low numbers. You can circle one of the numbers in the middle if that's how you feel. Got the idea? Any questions? OK. You can begin.

When each subject was finished, Experimeter 2 collected the material and unobtrusively recorded the time. The average time for completing the task was 16 minutes.

Cloze-scoring method. Children received cloze scores based on the number of correct words supplied for each high- and low-interest passage. Only the exact word was accepted as a correct response. The acceptance of synonyms does not increase the validity of the cloze procedure (Bormuth, 1965), and it decreases scoring objectivity and efficiency. Responses were considered to be correct despite spelling errors if the supplied word was clearly recognizable as the deleted word.

Results

Reading Achievement Test

The Scholastic Testing Service test had been administered by the school system one month prior to the study. Data were available for all but two students. The reading comprehension score for boys was 22.54 and for girls was 28.68; this difference is statistically significant ($t = 7.14$, $df = 83$, $p < .01$). Achievement test scores also were correlated with the total number of deletions correctly supplied on the cloze test. The relationship between the achievement and cloze measures was significant ($r = .69$, $df = 83$, $p < .01$). This is consistent with previous evidence that the cloze procedure is a valid measure of reading comprehension.

Picture Ratings

Since the purpose of the study was to compare boys and girls on high- and low-interest material, it was important to ensure that the highest three and lowest three picture ratings were similar for each sex. For each child, the highest possible combined rating for his or her three most interesting pictures is 21. The lowest possible combined rating for his or her three least interesting pictures is 3. For boys, the average combined rating for the three most preferred pictures was 20.90 and for girls it was 20.95 ($t = .76$, $df = 85$). The average combined rating on the three lowest rated pictures was 3.96 for boys and 4.15 for girls ($t = .45$, $df = 85$). Thus, the pictures provided very interesting and very uninteresting topics for individuals of both sexes.

Preference Ratings

If the picture assessment technique is valid, it should predict children's reading preferences. Table 1 presents the data on children's desire to read more about high- and low-interest topics after having read all six passages. Given three 1–7 rating scales for the high-interest material and three for low-interest material, the reading preference scores could range from 3 to 21 for each level of interest.

TABLE 1

Reading Preference Ratings

Interest level	Sex of student			
	Boys		Girls	
	M	*SD*	*M*	*SD*
High	15.63	4.84	14.66	4.59
Low	7.22	4.20	7.84	4.10

TABLE 2
CLOZE READING SCORES

Interest level	Sex of student			
	Boys		Girls	
	M	*SD*	*M*	*SD*
High	9.33	5.32	10.71	4.48
Low	5.57	3.91	9.32	4.49

A 2 × 2 (Sex × Interest) analysis of variance was performed on these data. Results indicated that children significantly preferred the high-interest material ($F = 206.37$, $df = 1/85$, $p < .01$), and there was no significant difference between boys' and girls' ratings ($F = .05$, $df = 1/85$). Furthermore, the interaction of sex and interest was nonsignificant ($F = 2.16$, $df = 1/85$). Both boys and girls, then, expressed significantly greater preference for reading the material that corresponded to their high-interest areas. These results validate the use of the picture-interest assessment method since picture ratings predicted the reading preferences for both sexes.

Cloze Scores

Table 2 presents the reading comprehension data. Each child received high- and low-interest cloze scores based on the number of his or her correct responses. Given three high- and three low-interest paragraphs and 10 deletions per paragraph, the high- and low-interest cloze scores could each range from 0 to 30. A 2 × 2 (Sex × Interest) analysis of variance indicated that girls read significantly better than boys ($F = 8.33$, $df = 1/85$, $p < .01$) and children comprehended more of the high- than low-interest paragraphs ($F = 38.81$, $df = 1/85$). Both of these main effects, however, are qualified by a significant Sex × Interest interaction ($F = 7.17$, $df = 1/85$, $p < .01$).

As Table 2 shows, the effect of interest was stronger for boys than for girls. Tukey's honestly significant difference test (Kirk, 1968) was used for comparisons between groups. Boys' high-interest performance was significantly superior to their low-interest performance ($p < .01$). Girls' performance, however, was not significantly affected by the interest level of the material. The sex differences in reading comprehension also varied with the interest level of passages. On low-interest paragraphs, boys scored significantly below girls ($p < .01$). However, the two groups were not significantly different on high-interest materials. Sex differences in reading performance, then, depended on children's interest in the material.

One interpretation of the interaction of sex and interest level is that the effect of interest is stronger for children who are lower achieving. Since boys in the present study scored lower than girls, the relevant sex-related variable may be achievement level. To examine this interpretation, the sample was divided at the median based on the Scholastic Testing Service reading comprehension score. Table 3 presents data on the effects of interest on high- and low-achieving students. The effect of interest on reading comprehension is similar in both high- and low-achieving groups. A two-way (Achievement Level × Interest) analysis of variance indicated that high-achieving children comprehended more than low-achieving children ($F = 29.90$, $df = 1/83$, $p < .01$) and that comprehension was greater on high-interest than low-interest materials ($F = 35.79$, $df = 1/83$, $p < .01$). Most important, the interaction of Achievement Level × Interest was nonsignificant ($F = .00$, $df = 1/83$). Thus, the Sex × Interest interaction obtained in the preceding analysis is not a function of differential impact of interest on high- versus low-achieving students. The reading comprehension of both high- and low-achieving students was facilitated by high-interest material.

TABLE 3
CLOZE READING SCORES BY ACHIEVEMENT LEVEL

Interest level	Achievement level			
	High achievers		Low achievers	
	M	*SD*	*M*	*SD*
High	12.09	3.91	7.32	4.72
Low	9.34	3.97	4.73	3.84

In summary, both girls and boys preferred to read more about the material corresponding to their high-interest topics. There was also a significant effect of interest on reading comprehension. However, the effect of interest on comprehension was stronger for boys than girls. The sex difference in reading performance was significant on low-interest material but not on high-interest material. Finally, the effect of interest was similar for high- and low-achieving students.

Discussion

Results of this study indicate that previous generalizations about the inferior reading performance of boys need qualification. When boys were interested in the material, they read as well as girls, while lack of interest produced results similar to those reported in earlier studies. Boys' performance appears to have been particularly facilitated by the high-interest material, while girls comprehended nearly as much of the low-interest material as the high. From these findings, it appears that when boys read more poorly than girls, it was because of low motivation rather than lack of skill. Since boys read as well as girls on the high-interest material, it indicates that they had the ability to comprehend material if they were motivated.

What is needed is a motivational account which explains why boys were facilitated by the high-interest material and girls were relatively unaffected by the interest level of material. One potential explanation is based on theory and data concerning the relationship between children's sex role standards and their reading performance. Kagan (1964) has noted a discrepancy between boys' sex role standards and the sex role connotations of school and reading. There is evidence that boys and girls perceive reading as a feminine activity (Kagan, 1964; Stein & Smithells, 1969). If reading is sex-appropriate for girls, they may be likely to read well regardless of the interest level of the task. If reading is sex-inappropriate for boys, they may require the additional incentive provided by high-interest material to read well. The three pictures most interesting to boys were race cars, basketball players, and astronauts. Topics such as these may make reading more "masculine." Stein, Pohly, and Mueller (1971) have found that boys' performance is particularly affected by the sex-typing of an activity.

An alternative to this motivational explanation is the possibility that the interaction of sex and interest results from girls' greater familiarity with vocabulary in a wider range of topic areas. If girls read more widely than boys, as some have suggested (e.g., Furness, 1963), then they might perform similarly on high- and low-interest passages, while boys might perform better on high-interest material containing more familiar vocabulary. The notion that girls' wider reading activity extends to low-interest topics seems somewhat implausible in light of the current finding that girls, like boys, showed little enthusiasm for the low-interest material. Still, the familiarity-of-vocabulary interpretation merits consideration. It is currently being evaluated by controlling the vocabulary of passages while manipulating the topics. If the familiarity interpretation is correct, comprehension of vocabulary-controlled passages should be similar across interest level for both boys and girls.

Another area for further research is the effect of interest on reading comprehension at different levels of passage difficulty. The passages used in the present study were somewhat difficult. The Dale-Chall formula indicated that the average passage was about two grades above the grade placement of children in the study. Still, children supplied the correct response for 29% of the cloze deletions, a percentage similar to that obtained when fourth- and seventh-grade children are given material selected by teachers to be appropriate to their grade level (Hansen & Hesse, 1974). Other studies indicate that a cloze score of 29% corresponds to approximately 65% correct on a multiple-choice achievement test (Bormuth, 1967; Rankin & Culhane, 1969) and to considerable pretest to posttest information gain on a multiple-choice test (Bormuth, 1968). Thus, children appear to have been deriving information from reading the pas-

sages used in the present study. Future studies might systematically vary passage difficulty in order to study the influence of this factor.

A related issue is the effect of interest on sex differences in reading at different ages. Fifth-grade children were tested in the present study. Sex differences in reading achievement narrow between elementary and high school grades. Perhaps fifth-grade boys are transitional in their development and are more likely to be influenced by the interest level of the material than they would be at an earlier or later period.

Finally, the present study has implications for reading skill assessment and instruction. The fact that sex difference found on the school-administered achievement test and the low-interest cloze passages, but not on the high-interest cloze material, implies that achievement test performance occurs under low-interest conditions. Achievement tests may not provide the kind of material that encourages boys to demonstrate their abilities. An interesting and critical issue is whether high-interest material would have long-term instructional benefits as well as short-term assessment affects. Stanchfield (1967) has reported success in teaching reading with high-interest materials. However, in that study, the use of interesting material was one of a number of instructional interventions. Furthermore, reading material in that study was the same for all children rather than being based on individual interests. It remains to be determined whether the intervention of an individualized high-interest reading curriculum would narrow the gap between boys' and girls' reading performance. The interest assessment procedure used in the present study could easily be adapted to help individualize classroom reading programs. Results of the procedure predicted the reading preferences of boys and girls.

REFERENCES

Asher, S. R., & Gottman, J. M. Sex of teacher and student reading achievement. *Journal of Educational Psychology*, 1973, **65**, 168–171.

Bernstein, M. R. The relationship between interest and reading comprehension. *Journal of Educational Research*, 1955, **49**, 283–288.

Blom, G. E. Sex differences in reading disability. In E. Calkins (Ed.), *Reading forum*. Bethesda, Md.: National Institute of Neurological Disease and Stroke, 1971.

Blom, G. E., Waite, R. R., & Zimet, S. Content of first grade reading books. *The Reading Teacher*, 1968, **21**, 317–323.

Bormuth, J. R. Validities of grammatical and semantic classifications of cloze test scores. *Proceedings of the International Reading Association*, 1965, **10**, 283–286.

Bormuth, J. R. Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, 1967, **10**, 291–299.

Bormuth, J. R. Empirical determination of the instructional reading level. *Proceedings of the International Reading Association*, 1968, **13**, 716–721.

Britannica Junior Encyclopedia. Chicago: Encyclopedia Britannica, 1970.

Dale, E., & Chall, J. S. A formula for predicting readability: Instructions. *Educational Research Bulletin*, 1948, **27**, 37–54.

Furness, E. L. Researches on reading interests. *Education*, 1963, **84**, 3–7.

Hansen, L. H., & Hesse, K. D. An assessment of reading literacy. Madison, Wisconsin: Madison Public Schools, Department of Research and Development, 1974.

Kagan, J. The child's sex role classification of school objects. *Child Development*, 1964, **35**, 1051–1056.

Kirk, R. E. *Experimental design: Procedures for the behavioral sciences*. Belmont, Calif.: Brooks/Cole, 1968.

Klein, H. A. Interest and comprehension in sex-typed materials. Paper presented at the International Reading Association Conference, Kansas City, May 1969.

Neale, D. C., Gill, N., & Tismer, W. Relationship between attitudes toward school subjects and school achievement. *Journal of Educational Research*, 1970, **63**, 232–237.

Rankin, E. F., & Culhane, J. W. Comparable cloze and multiple-choice test scores. *Journal of Reading*, 1969, **13**, 193–198.

Shnayer, S. W. Some relationships between reading interests and reading comprehension. Unpublished doctoral dissertation, University of California, Berkeley, 1967.

Stanchfield, J. M. The effect of high-interest materials on reading achievement in the first grade. *National Reading Conference Yearbook*, 1967, **16**, 58–61.

Stein, A. H., Pohly, S. R., & Mueller, E. The influence of masculine, feminine, and neutral tasks on children's achievement behavior, expectancies of success, and attainment values. *Child Development*, 1971, **42**, 195–207.

Stein, A. H., & Smithells, J. Age and sex differences in children's sex-role standards about

achievement. *Developmental Psychology*, 1969, **1**, 252–259.

Stroud, J. B., & Lindquest, E. F. Sex differences in achievement in the elementary and secondary schools. *Journal of Educational Psychology*, 1942, **33**, 657–667.

Taylor, W. L. "Cloze procedure:" A new tool for measuring readability. *Journalism Quarterly*, 1953, **30**, 415–433.

Walsh, S. P. *General Encyclopedias in Print 1973–1974: A Comparative Analysis*. New York: R .R. Bowker, 1973.

(December 17, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 688-701

ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN RANDOMIZED AND NONRANDOMIZED STUDIES[1]

DONALD B. RUBIN[2]

Educational Testing Service, Princeton, New Jersey

A discussion of matching, randomization, random sampling, and other methods of controlling extraneous variation is presented. The objective is to specify the benefits of randomization in estimating causal effects of treatments. The basic conclusion is that randomization should be employed whenever possible but that the use of carefully controlled nonrandomized data to estimate causal effects is a reasonable and necessary procedure in many cases.

Recent psychological and educational literature has included extensive criticism of the use of nonrandomized studies to estimate causal effects of treatments (e.g., Campbell & Erlebacher, 1970). The implication in much of this literature is that only properly randomized experiments can lead to useful estimates of causal effects. If taken as applying to all fields of study, this position is untenable. Since the extensive use of randomized experiments is limited to the last half century,[3] and in fact is *not* used in much scientific investigation today,[4] one is led to the conclusion that most scientific "truths" have been established without using randomized experiments. In addition, most of us successfully determine the causal effects of many of our everyday actions, even interpersonal behaviors, without the benefit of randomization.

Even if the position that causal effects of treatments can only be well established from randomized experiments is taken as applying only to the social sciences in which there are currently few well-established causal relationships, its implication—to ignore existing observational data—may be counter-productive. Often the only immediately available data are observational (nonrandomized) and either (*a*) the cost of performing the equivalent randomized experiment to test all treatments is prohibitive (e.g., 100 reading programs under study); (*b*) there are ethical reasons why the treatments cannot be randomly assigned (e.g., estimating the effects of heroin addiction on intellectual functioning); or (*c*) estimates based on results of experiments would be delayed many years (e.g., effect of childhood intake of cholesterol on longevity). In cases such as these, it seems more reasonable to try to estimate the effects of the treatments from nonrandomized studies than to ignore these data and dream of the ideal experiment or make "armchair" decisions without the benefit of data analysis. Using the indications from nonrandomized studies, one can, if necessary, initiate randomized experiments for those treatments that require better estimates or that look most promising.

The position here is *not* that randomization is overused. On the contrary, given a choice between the data from a randomized experiment and an equivalent nonrandomized study, one should choose the data from the experiment, especially in the social sciences where much of the variability is often unassigned to particular causes. However, we will develop the position that nonrandomized studies as well as randomized

[1] I would like to thank E. J. Anastasio, A. E. Beaton, W. G. Cochran, K. M. Kazarow, and R. L. Linn for helpful comments on earlier versions of this paper. I would also like to thank the U. S. Office of Education for supporting work on this paper under contract OEC-0-71-3715.

[2] Requests for reprints should be sent to Donald B. Rubin, Division of Data Analysis Research, Educational Testing Service, Princeton, New Jersey 08540.

[3] Essentially since Fisher (1925).

[4] For example, in Davies (1954), a well-known textbook on experimental design in industrial work, randomization is not emphasized.

experiments can be useful in estimating causal treatment effects.

In order to avoid unnecessary complication, we will restrict discussion to the very simple study consisting of $2N$ units (e.g., subjects), half having been exposed to an experimental (E) treatment (e.g., a compensatory reading program) and the other half having been exposed to a control (C) treatment (e.g., a regular reading program). If Treatments E and C were assigned to the $2N$ units randomly, that is, using some mechanism that assured each unit was equally likely to be exposed to E as to C, then the study is called a randomized experiment or more simply an experiment; otherwise, the study is called a nonrandomized study, a quasi-experiment, or an observational study. The objective is to determine for some population of units (e.g., underprivileged sixth-grade children) the "typical" causal effect of the E versus C treatment on a dependent Variable Y, where Y could be dichotomous (e.g., success–failure) or more continuous (e.g., score on a given reading test). The central question concerns the benefits of randomization in determining the causal effect of the E versus C treatment on Y.

Defining the Causal Effect of the E versus C Treatment

Intuitively, the causal effect of one treatment, E, over another, C, for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to E initiated at t_1 and what would have happened at t_2 if the unit had been exposed to C initiated at t_1: "If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone," or "Because an hour ago I took two aspirins instead of just a glass of water, my headache is now gone." Our definition of the causal effect of the E versus C treatment will reflect this intuitive meaning.

First define a trial to be a unit and an associated pair of times, t_1 and t_2, where t_1 denotes the time of initiation of a treatment and t_2 denotes the time of measurement of a dependent variable, Y, where $t_1 < t_2$. We restrict our attention to Treatments E and C that could be randomly assigned; thus, we assume (*a*) a time of initiation of treatment can be ascertained for each unit exposed to E or C and (*b*) E and C are exclusive of each other in the sense, that a trial cannot simultaneously be an E trial and a C trial (i.e., if E is defined to be C plus some action, the initiation of both is the initiation of E; if E and C are alternative actions, the initiation of both E and C is the initiation of neither of these but rather of a third treatment, E + C).

Now define the causal effect of the E versus C treatment on Y for a particular trial (i.e., a particular unit and associated times t_1, t_2) as follows:

> Let y(E) be the value of Y measured[5] at t_2 on the unit, given that the unit received the experimental Treatment E initiated at t_1;
>
> Let y(C) be the value of Y measured at t_2 on the unit given that the unit received the control Treatment C initiated at t_1;
>
> Then y(E) − y(C) is the causal effect of the E versus C treatment on Y for that trial, that is, for that particular unit and the times t_1, t_2.

For example, assume that the unit is a particular child, the experimental treatment is an enriched reading program, and the control treatment is a regular reading program. Suppose that if the child were given the enriched program initiated at time

[5] The measured value of Y stated with reference to time t_2 is considered the "true" value of Y at t_2. This position can be justified by defining Y by a measuring instrument that always yields the measured Y (e.g., Y is the score on a particular IQ test as recorded by the subject's teacher). Since an "error" in the measured Y can only be detected by a "better" measuring instrument (e.g., a machine-produced score on that same IQ test), the values of a "truer" score can be viewed as the values of a different dependent variable. Clearly, any study is more meaningful to the investigator if the dependent variable better reflects underlying concepts he feels are important (e.g., is more accurate) but that does not imply he must consider errors about some unmeasurable "true score." For the reader who prefers the concept of such errors of measurement, he may consider the following discussion to assume negligible "technical errors" so that Y is essentially the "true" Y

t_1, 10 days later at time t_2 he would have a score of 38 items correct on a reading test; and suppose that if the child instead were given the regular program initiated at time t_1, at time t_2 he would score 34 items correct. Then the causal effect on the reading test for that trial (that child and times t_1, t_2) of the enriched program versus the regular program is $38 - 34 = 4$ more items correct.

The problem in measuring $y(E) - y(C)$ is that we can never observe both $y(E)$ and $y(C)$ since we cannot return to time t_1 to give the other treatment. We may have the same unit measured on both treatments in two trials (a repeated measure design), but since there may exist carryover effects (e.g., the effect of the first treatment wears off slowly) or general time trends (e.g., as the child ages, his learning ability increases), we cannot be certain that the unit's responses would be identical at both times.

Assume now that there are M trials for which we want the "typical" causal effect. For simplicity of exposition, assume that each trial is associated with a different unit and expand the above notation by adding the subscript j to denote the jth trial ($j = 1, 2, \cdots, M$); thus $y_j(E) - y_j(C)$ is the causal effect of the E versus C treatment for the jth trial, that is, the jth unit and the associated times of initiation of treatment, t_{1j}, and measurement of Y, t_{2j}.

An obvious definition of the "typical" causal effect of the E versus C treatment for the M trials is the average (mean) causal effect for the M trials:

$$\frac{1}{M}\sum_{j=1}^{M}[y_j(E) - y_j(C)].$$

Even though other definitions of typical are interesting,[6] they lead to more complications when discussing properties of estimates under randomization. Hence we assume the average causal effect is the desired typical causal effect for the M trials and proceed to the problem of its estimation given the obvious constraint that we can never actually measure both $y_j(E)$ and $y_j(C)$ for any trial.

Randomization, Matching, and Estimating the Typical Causal Effect in the $2N$ Trial Study

For now assume that the objective is to estimate the typical causal effect only for the $2N$ trials in the study. Of course, in order for the results of a study to be of much interest, we must be able to generalize to units and associated times other than those in the study. However, the issue of generalizing results to other trials is discussed separately from the issue of estimating the typical causal effect for the trials under study. Also, for now we only consider the simple and standard estimate of the typical causal effect of E versus C: the average Y difference between those units who received E and those units who received C. After considering this estimate when there are only two trials in the study and then when there are $2N$ ($N > 1$) trials in the study, we will more formally discuss two benefits of randomization.

Two-Trial Study

Suppose there are two trials under study, one trial having a unit exposed to E and the other having a unit exposed to C. The typical causal effect for the two trials is

$$\tfrac{1}{2}[y_1(E) - y_1(C) + y_2(E) - y_2(C)]. \quad [1]$$

The estimate of this quantity from the study, the difference between the measured Y for the unit who received E and the measured Y for the unit who received C, is either

$$y_1(E) - y_2(C) \quad [2]$$

or

$$y_2(E) - y_1(C) \quad [3]$$

[6] Notice that if all but one of the individual causal effects are small and that one is very large, the average causal effect may be substantially larger than all but one of the individual causal effects and thus not very "typical." Other possible definitions of the typical causal effects for the M trials are the median causal effect (the median of the individual causal effects) or the midmean causal effect (the average of the middle half of the individual causal effects). If the individual causal effects, $y_j(E) - y_j(C)$, are approximately symmetrically distributed about a central value, sensible definitions of "typical" will yield similar values.

depending upon which unit was assigned E. Neither Equation 2 nor Equation 3 is necessarily close to Equation 1 or to the causal effect for either unit

$$y_1(E) - y_1(C) \quad [4]$$

or

$$y_2(E) - y_2(C), \quad [5]$$

even if these individual causal effects are equal. If the Treatments E and C were randomly assigned to units, we are equally likely to have observed the difference in Equation 2 as that in Equation 3, so that the average or "expected" difference in Y between experimental and control units is the average of Equations 2 and 3,

$$\tfrac{1}{2}\,[y_1(E) - y_2(C)] + \tfrac{1}{2}\,[y_2(E) - y_1(C)]$$

which equals Equation 1, the typical causal effect for the two trials. For this reason, if the treatments are randomly assigned, the difference in Y between the experimental and control units is called an "unbiased" estimate of the desired typical causal effect.

Now suppose that the two units are very similar in the way they respond to the E and C treatments at the times of their trials. By this we mean that on the basis of "extra information," we know $y_1(E)$ is about equal to $y_2(E)$ and $y_1(C)$ is about equal to $y_2(C)$; that is, the two trials are closely "matched" with respect to the effects of the two treatments. It then follows that Equation 2 is about equal to Equation 3, and both are about equal to the desired typical causal effect in Equation 1. In fact, if the two units react identically in their trials, Equation 5 = Equation 4 = Equation 3 = Equation 2 = Equation 1, and randomization is absolutely irrevelant. Clearly, having closely "matched" trials increases the closeness of the calculated experimental minus control difference to the typical causal effect for the two trials, while random assignment of treatments does *not* improve that estimate.

Although two-trial studies are almost unheard of in the behavioral sciences, they are not uncommon in the physical sciences. For example, when comparing the heat expansion rates (per hour) of a metal alloy in oxygen and nitrogen, an investigator might use 2 one-foot lengths of the alloy. Because the lengths of alloy are so closely matched before being exposed to the treatment (almost identical compositions and dimensions), the units should respond almost identically to the treatments even when initiated at different times, and thus the calculated experimental (oxygen) minus control (nitrogen) difference should be an excellent estimate of the typical causal effect, Equation 1.

A skeptical observer, however, could always claim that the experimental minus control difference is not a good estimate of the typical causal effect of the E versus C treatment because the two units were not absolutely identical prior to the application of the treatments. For example, he could claim that the length of alloy molded first would expand more rapidly. Hence, he might argue that what was measured was really the effect of the difference in order of manufacture, not the causal effect of the oxygen versus nitrogen treatment. Since units are never absolutely identical before the application of treatments, this kind of argument, whether "sensible" or not, can always be made. Nevertheless, if the two trials are closely matched with respect to the expected effects of the treatments, that is, if (*a*) the two units are matched prior to the initiation of treatments on all variables thought to be important in the sense that they causally affect Y and (*b*) the possible effect of different times of initiation of treatment and measurement of Y are controlled, then the investigator can be confident that he is in fact measuring the causal effect of the E versus C treatment for those two trials. This kind of confidence is much easier to generate in the physical sciences where there are models that successfully assign most variability to specific causes than in the social sciences where often important causal variables have not been identified.

Another source of confidence that the experimental minus control difference is a good estimate of the causal effect of E versus C is replication: Are similar results obtained under similar conditions? One type of replication is the inclusion of more than two trials in the study.

The 2N Trial Study

Suppose there are $2N$ trials ($N > 1$) in the study, half with N units having received the E treatment and the other half with N units having received the C treatment. The immediate objective is to find the typical causal effect of the E versus C treatment on Y for the $2N$ trials, say τ:

$$\tau = \frac{1}{2N} \sum_{j=1}^{2N} [y_j(E) - y_j(C)].$$

Let S_E denote the set of indices of the E trials and S_C denote the set of indices of the C trials ($S_E \cup S_C = \{i = 1, 2, \cdots, 2N\}$). Then the difference between the average observed Y in the E trials and the average observed Y in the C trials can be expressed as

$$\bar{y}_d = \frac{1}{N} \sum_{j \in S_E} y_j(E) - \frac{1}{N} \sum_{j \in S_C} y_j(C),$$

where $\sum_{j \in S_E}$ and $\sum_{j \in S_C}$ indicate, respectively, summation over all indices in S_E (i.e., all E trials) and over all indices in S_C (i.e., all C trials). We now consider how close this estimate $\bar{y}_d$ is to the typical causal effect τ and what advantage there might be if we knew the treatments were randomly assigned.

First, assume that for each unit receiving E there is a unit receiving C, and the two units react identically at the times of their trials; that is, the $2N$ trials are actually N perfectly matched pairs. We now show that the estimate $\bar{y}_d$ in this case equals τ. $\bar{y}_d$ can be expressed as the average experimental minus control (E − C) difference across the N matched trials. Since the (E − C) difference in each matched pair of trials is the typical causal effect for both trials of that pair, the average of those differences is the typical causal effect for all N pairs and thus all $2N$ trials. This result holds whether the treatments were randomly assigned or not. In fact, if one had N identically matched pairs, a "thoughtless" random assignment could be worse than a nonrandom assignment of E to one member of the pair and C to the other. By "thoughtless" we mean some random assignment that does not assure that the members of each matched pair get different treatments—picking the N indices to receive E "from a hat" containing the numbers 1 through $2N$ rather than tossing a fair coin for each matched pair to see which unit is to receive E.

In practice, of course, we never have exactly matched trials. However, if matched pairs of trials are very similar in the sense that prior to the initiation of treatments the investigator has controlled those variables that might appreciably affect Y, then $\bar{y}_d$ should be close to τ. If, in addition, the estimated causal effect is replicable in the sense that the N individual estimated causal effects for each matched pair are very similar, the investigator might feel even more confident that he is in fact estimating the typical causal effect for the $2N$ trials (e.g., $2N$ children from the same school matched by sex and initial reading score into N pairs, with the same observed E − C difference in final score in each matched pair). Similarly, if the trials are not pair-matched but are all similar (e.g., all children are males from the same school with similar pretest scores) and if we observe that all $y_j(E)$ $j \in S_E$ are about equal and all $y_j(C)$ $j \in S_C$ are about equal, the investigator would also feel confident that he is in fact estimating the typical causal effect for the $2N$ trials.

Nevertheless, it is obvious that if treatments were systematically assigned to units, the addition of replication evidence cannot dissuade the critic who believes the effect being measured is due to a variable used to assign treatments (e.g., in the reading study, if more active children always received the enriched program, or in the heat-expansion study, if the first molded alloy was always measured in oxygen). If treatments were randomly assigned, all systematic sources of bias would be made random, and thus it would be unlikely, especially if N is large, that almost all E trials would be with the more active children or the first molded alloy. Hence, any effect of that variable would be at least partially balanced in the sense of systematically favoring neither the E treatment nor the C treatment over the $2N$ trials. In addition, using the replications, there could be evidence to refute the skeptic's claim of the importance of that variable (e.g., in each matched trial we get about the

same estimate whether the more active child gets E or C). Of course, if we knew the skeptic's claim beforehand, a specific control of this additional variable would be more advisable than relying on randomization (e.g., in a random half of the matched trials assign E to the more active child, and in the other half assign C to the more active child, or include the child's activity as a matching variable).

It is important to realize, however, that whether treatments are randomly assigned or not, no matter how carefully matched the trials, and no matter how large N, a skeptical observer could always eventually find some variable that systematically differs in the E trials and C trials (e.g., length of longest hair on the child) and claim that $\bar{y}_d$ estimates the effect of this variable rather than τ, the causal effect of the E versus C treatment. Within the experiment there can be no refutation of this claim; only a logical argument explaining that the variable cannot causally affect the dependent variable or additional data outside the study can be used to counter it.

Two Formal Benefits of Randomization

If randomization can never assure us that we are correctly estimating the causal effect of E versus C for the $2N$ trials under study, what are the benefits of randomization besides the intuitive ones that follow from making all systematic sources of bias into random ones? Formally, randomization provides a mechanism to derive probabilistic properties of estimates without making further assumptions. We will consider two such properties that are important:

1. The average E − C difference is an "unbiased" estimate of τ, the typical causal effect for the $2N$ trials.
2. Precise probabilistic statements can be made indicating how unusual the observed E − C difference, $\bar{y}_d$, would be under specific hypothesized causal effects.

More advanced discussion of the formal benefits of randomization may be found in Sheffé (1959) and Kempthorne (1952).

Unbiased Estimation over the Randomization Set

We begin by defining the "randomization set" to be the set of r allocations that were equally likely to be observed given the randomization plan. For example, if the treatments were randomly assigned to trials with no restrictions (the completely randomized experiment, Cochran & Cox, 1957), each one of the $\binom{2N}{N}$ possible allocations of N trials to E and N trials to C was equally likely to be the observed allocation. Thus, the collection of all of these $r = \binom{2N}{N}$ allocations is known as the randomization set for this completely randomized experiment. If the treatments were assigned randomly within matched pairs (the randomized blocks experiment, Cochran & Cox, 1957), any of the 2^N allocations, with each member of the pair receiving a different treatment, was equally likely to be the observed one. Hence, for the experiment with randomization done within matched pairs, the collection of these $r = 2^N$ equally likely allocations is known as the randomization set.

For each of the r possible allocations in the randomization set, there is a corresponding average E − C difference that would have been calculated had that allocation been chosen. If the expectation (i.e., average) of these r possible average differences equals τ, the average E − C difference is called unbiased over the randomization set for estimating τ. We now show that given randomly assigned treatments, the average E − C difference is an unbiased estimate of τ, the typical causal effect for the $2N$ trials.

By the definition of random assignment, each trial is equally likely to be an E trial as a C trial. Hence, the contribution of the jth trial ($j = 1, \cdots, 2N$) to the average E − C difference in half of the r allocations in the randomization set is $y_j(E)/N$ and in the other half is $-y_j(C)/N$. The expected contribution of the jth trial to the average E − C difference is therefore

$$\tfrac{1}{2}[y_j(E)/N] + \tfrac{1}{2}[-y_j(C)/N].$$

Summing over all $2N$ trials we have, the expectation of the average E − C difference over the r allocations in the randomization set is

$$\frac{1}{2N}\sum_{j=1}^{2N}[y_j(E) - y_j(C)],$$

which is the typical causal effect for the $2N$ trials, τ.

Although the unbiasedness of the E − C difference is appealing in the sense that it indicates that we are tending to estimate τ, its impact is not immediately overwhelming: the one E − C difference we have observed, $\bar{y}_d$, may or may not be close to τ. In a vague sense we may believe $\bar{y}_d$ should be close to τ because the unbiasedness indicates that "on the average" the E − C difference is τ, but this belief may be tempered when other properties of the estimate are revealed; for example without additional constraints on the symmetry of effects, the average E − C difference is *not* equally likely to be above τ as below it.

In addition, after observing the values of some important unmatched variable, we may no longer believe $\bar{y}_d$ tends to estimate τ. For example, suppose in the study of reading programs, initial reading score is not a matching variable, and after the experiment is complete we find that the average initial score for the children exposed to E was higher than for those exposed to C. Clearly we would now believe that $\bar{y}_d$ probably overestimates τ even if treatments were randomly assigned.

In sum, the unbiasedness of the E − C difference for τ follows from the random assignment of treatments; it is a desirable property because it indicates that "on the average" we tend to estimate the correct quantity, but it hardly solves the problem of estimating the typical causal effect. As yet we have no indication whether to believe $\bar{y}_d$ is close to τ nor to any ability to adjust for important information we may possess.

Probabilistic Statements from the Randomization Set

A second formal advantage of randomization is that it provides a mechanism for making precise probabilistic statements indicating how unusual the observed E − C difference, $\bar{y}_d$, would be under specific hypotheses. Suppose that the investigator hypothesizes exactly what the individual causal effects are for each of the $2N$ trials and these hypothesized values are $\tilde{\tau}_j$, $j = 1, \cdots, 2N$. The hypothesized typical causal effect for the $2N$ trials is thus

$$\tilde{\tau} = \frac{1}{2N}\sum_{j=1}^{2N}\tilde{\tau}_j.$$

Having the $\tilde{\tau}_j$ and the observed $y_j(E)$, $j \epsilon S_E$ and $y_j(C)$, $j \epsilon S_C$, we can calculate hypothesized values, say $\tilde{y}_j(C)$ and $\tilde{y}_j(E)$, for all of the $2N$ trials. For $j \epsilon S_E$, $y_j(E)$ is observed and $y_j(C)$ is unobserved; hence, for these trials $\tilde{y}_j(E) = y_j(E)$ and $\tilde{y}_j(C) = y_j(E) - \tilde{\tau}_j$. For $j \epsilon S_C$, $y_j(C)$ is observed and $y_j(E)$ is unobserved; hence, for these trials $\tilde{y}_j(C) = y_j(C)$ and $\tilde{y}_j(E) = y_j(C) + \tilde{\tau}_j$. Thus, we can calculate hypothesized $\tilde{y}_j(E)$ and $\tilde{y}_j(C)$ for all $2N$ trials, and using these, we can calculate an hypothesized average E − C difference for each of the r allocations of the $2N$ trials in the randomization set.

Suppose that we calculate the r hypothesized average E − C differences and list them from high to low, noting which E − C difference corresponds to the S_E, S_C allocation we have actually observed. This difference, $\bar{y}_d$, is the only one which does not use the hypothesized $\tilde{\tau}_j$. If treatments were assigned completely at random to the trials and the hypothesized $\tilde{\tau}_j$ are correct, any one of the $r = \binom{2N}{N}$ differences was equally likely to be the observed one; similarly, if treatments were randomly assigned within matched pairs, each of the $r = 2^N$ differences with each member of a matched pair getting a different treatment was equally likely to be the observed one. Intuitively, if the hypothesized $\tilde{\tau}_j$ are essentially correct, we would expect the observed difference $\bar{y}_d$ to be rather typical of the (r − 1) other differences that were equally likely to be observed; that is, $\bar{y}_d$ should be near the center of the distribution of the r E − C differences. If the observed difference is in the tail of distribution and therefore not typical of the r differences, we might doubt the correctness of the hypothesized $\tilde{\tau}_j$.

Since the average of the r E − C differences is the hypothesized typical causal

effect, $\bar{\tau}$, and the r allocations are equally likely, we can make the following probabilistic statement:

> Under the hypothesis that the causal effects are given by the τ_j, $j = 1, \cdots, 2N$, the probability that we would observe an average E − C difference that is as far or farther from $\bar{\tau}$ than the one we have observed is m/r where m is the number of allocations in the randomization set that yield E − C differences that are as far or farther from $\bar{\tau}$ than $\bar{y}_d$.

If this probability, called the "significance level" for the hypothesized $\bar{\tau}_j$, is very small, we either must admit that the observed value is unusual in the sense that it is in the tail of the distribution of the equally likely differences, or we must reject the plausibility of the hypothesized $\bar{\tau}_j$.

The most common hypothesis for which a significance level is calculated is that the E versus C treatment has no effect on Y whatsoever (i.e., $\bar{\tau}_j \equiv 0$). Other common hypotheses assume that the effect of the E versus C treatment on Y is a nonzero constant (i.e., $\bar{\tau}_j \equiv \bar{\tau}_0$) for all trials.[7]

The ability to make precise probabilistic statements about the observed $\bar{y}_d$ under various hypotheses without additional assumptions is a tremendous benefit of randomization especially since $\bar{y}_d$ tends to estimate τ. However, one must realize that these simple probabilistic statements refer only to the $2N$ trials used in the study and do not reflect additional information (i.e., other variables) that we may also have measured.

Presenting the Results of an Experiment as Being of General Interest

Before presenting the results of an experiment as being relevant, an investigator should believe that he has measured the causal effect of the E versus C treatment and not the effect of some extraneous variable. Also, he should believe that the result is applicable to a population of trials besides the $2N$ in the experiment.

Considering Additional Variables

As indicated previously, the investigator should be prepared to consider the possible effect of other variables besides those explicit in the experiment. Often additional variables will be ones that the investigator considers relevant because they may causally affect Y; therefore, he may want to adjust the estimate $\bar{y}_d$ and significance levels of hypotheses to reflect the values of these variables in the study. At times the variables will be ones which cannot causally affect Y even though in the study they may be correlated with the observed values of Y. An investigator who refuses to consider any additional variables is in fact saying that he does not care if $\bar{y}_d$ is a bad estimate of the typical causal effect of the E versus C treatment but instead is satisfied with mathematical properties (i.e., unbiasedness) of the process by which he calculated it.

Consider first the case of an obviously important variable. As an example, suppose in the reading study, with programs randomly assigned, we found that the average E − C difference in final score was four items correct and that under the hypothesis of no effects the significance level was .01; also assume that initial score was not a matching variable and in fact the difference in initial score was also four items correct. Admittedly, this is probably a rare event given the randomization, but rare events do happen rarely. Given that it did happen, we would indeed be foolish to believe $\bar{y}_d = 4$ items is a good estimate of τ and/or the implausibility of the hypothesis of no treatment effects indicated by the .01 significance level. Rather, it would seem more sensible to believe that $\bar{y}_d$ overestimates τ and significance levels underestimate the plausibility of hypotheses that suggest zero or negative values for τ.

A commonly used and obvious correction is to calculate the average E − C difference in gain score rather than final score. That is, for each trial there is a "pretest" score which was measured before the initiation of

---

[7] These hypotheses for a constant effect can be used to form "confidence limits" for τ. Given that the τ_j are constant, the set of all hypothesized $\bar{\tau}_0$ such that the associated significance level is greater than or equal to $\alpha = m/r$ form a $(1 - \alpha)$ confidence interval for τ: of the r such $(1 - \alpha)$ confidence intervals one could have constructed (one for each of the r allocations in the randomization set), $r(1 - \alpha) = r - m$ of them include the true value of τ assuming all $\tau_j = \tau$. See Lehmann (1959, p. 59) for the proof.

treatments, and the gain score for each trial is the final score minus the pretest score. More generally we will speak of a "prior" score or "prior" variable which would have the same value, x_j, whether the jth unit received E or C. It then follows given random assignment of treatments that the adjusted estimate (e.g., gain score)

$$\frac{1}{N}\sum_{j\in S_E}[y_j(E) - x_j] - \frac{1}{N}\sum_{j\in S_C}[y_j(C) - x_j]$$

remains an unbiased estimate of τ over the randomization set: Each prior score appears in half of the equally likely allocations as x_j/N and the other half as $-x_j/N$; hence, averaged over all allocations, the jth prior score has no effect.[8] But this result holds for any set of prior scores x_j, $j = 1, \cdots, 2N$, whether sensible or not. For example, in an experiment evaluating a compensatory reading program, with Y being the final score on a reading test, the prior variable "pretest reading score" or perhaps "IQ" properly scaled makes sense but "height in millimeters" does not. Also, why not use the prior variable "one half pretest score?"

Clearly, in order to make an intelligent adjustment for extra information, we cannot be guided solely by the concept of unbiasedness over the randomization set. We need some model for the effect of prior variables in order to use their values in an intelligent manner. For example, if the final score typically would equal the initial score if there were no E − C treatment effect (as with the length of the alloys in the heat expansion experiment), the gain score is perfectly reasonable. In the physical sciences, more complex models representing generally accepted functional relationships are often used; however, in the social sciences there are rarely such accepted relationships upon which to rely. What then does the investigator do in order to adjust intelligently the final reading scores for the subjects' varying IQs, grade levels, socioeconomic status, and so on? Apparently, he must be willing to make some assumptions about the functional form of the causal effect of these other variables on Y. If he assumes, perhaps based on indications in previous data, some "known" function for x_j (e.g., in the compensatory reading program example, suppose x_j equals $[.01 \times IQ]^2 \times$ pretest $\times$ [percentile of family income]), so that x_j is the same whether the jth unit received E or C, from the previous discussion the average E − C difference in adjusted scores remains an unbiased estimate of τ. If the investigator assumes a model whose parameters are unknown and estimates these parameters by some method from the data, in general the average E − C difference in adjusted scores is no longer unbiased over the randomization set because the adjustment for the jth trial depends on which trials received E and which received C (e.g., in the analysis of covariance, the estimated regression coefficients in general vary over the r allocations in the randomization set). Hence, forming an intelligent adjusted estimate may not be simple even in a randomized experiment.

Significance levels for any adjusted estimate can be found by calculating the adjusted estimate rather than the simple E − C difference for each of the equally likely allocations in the randomization set. However, if the adjusted estimate does not tend to estimate τ in a sensible manner, the resulting significance level may not be of much interest.

Now consider a variable that is brought to the investigator's attention, but he feels it cannot causally affect Y (e.g., in the compensatory reading example, age of oldest living relative). Eventually a skeptic can find such a variable that systematically differs in the E trials and the C trials even in the best of experiments. Considering only that variable, it is indeed unlikely given randomization that there would be such discrepancy between its values in E trials and C trials, but its occurrence cannot be denied. If the skeptic adjusts $\bar{y}_d$ by using a standard model (e.g., covariance), the adjusted estimate and related significance levels may then give misleading results (e.g., zero estimate of τ, hypothesis that

[8] If the prior score could vary depending on whether the unit received E or C (i.e., it is a variable measured after the initiation of the treatment), we would have no assurance that the adjusted E − C difference is an unbiased estimate over the randomization set.

all causal effects are zero, $\tau_j \equiv 0$, is very plausible). In fact, using such models one can obtain any estimated causal effect desired by searching for and finding a prior variable or combination of prior variables that yield the desired result. Such a search should be more difficult given that randomization was performed, but even with randomized data the investigator must be prepared to ignore variables that he feels cannot causally affect Y. On the other hand, he may want to adjust for such a variable if he feels it is a surrogate for an unmeasured variable that can causally affect Y (e.g., age of oldest living relative is a surrogate for mental stability of the family in the compensatory reading example).

The point of this discussion is that when trying to estimate the typical causal effect in the $2N$ trial experiment, handling additional variables may not be trivial without a well-developed causal model that will properly adjust for those prior variables that causally affect Y and ignore other variables that do not causally affect Y even if they are highly correlated with the observed values of Y. Without such a model, the investigator must be prepared to ignore some variables he feels cannot causally affect Y and use a somewhat arbitrary model to adjust for those variables he feels are important. An example which demonstrates that it is not always simple to interpret significant results in a randomized experiment with many prior variables recorded is the recent controversy over the utility of oral-diabetic drugs.[9]

Generalizing Results to Other Trials

In order to believe that the results of an experiment are of practical interest, we generally must believe that the $2N$ trials in the study are representative of a population of other future trials. For example, if the experimental treatment is a compensatory reading program and the trials are composed of sixth-grade school children with treatments initiated in fall 1970 and Y measured in spring 1971, the results are of little interest unless we believe they tell us something about future sixth graders who might be exposed to this compensatory reading program.

For simplicity, assume the $2N$ trials in the study are a simple random sample from a "target population" of M trials to which we want to generalize the results; by simple random sample we mean that each of the $\binom{M}{2N}$ ways of choosing the $2N$ trials is equally likely to be selected. If T is the typical (average) causal effect for all M trials, it then follows given random assignment of treatments that the average E − C difference for the $2N$ trials used is an unbiased estimate of T over the random sampling plan and over the randomization set. In other words, in each of the $r \times \binom{M}{2N}$ ways of choosing $2N$ trials from M trials and then randomly assigning N trials to E and N trials to C, there is a calculated average E − C difference, and the average of these $r \times \binom{M}{2N}$ differences is T: Because of the randomization and random sampling, each trial is equally likely to be an E trial as a C trial and thus contributes $y_j(E)/N$ to the E − C difference as often as it contributes $-y_j(C)/N$. It also follows that under a hypothesized set of causal effects, τ_j, $j = 1, \cdots, M$, the significance level (the probability that we would observe a difference as large as or larger than $\bar{y}_d$), given that we have sampled the $2N$ trials in the study, is m/r where m is the number of allocations in the randomization set that yield estimates as far or farther from τ than $\bar{y}_d$.[10]

If we let M grow to infinity (a reasonable assumption in many experiments when the population to which we want to generalize results is essentially unlimited, for example, all future sixth-grade students), some additional probabilistic results follow. For example, the usual covariance adjusted estimate is an unbiased estimate of T (not necessarily τ) over the random sampling plan and the randomization set, but whether the adjustment actually adjusts for the

[9] See for example Schor (1971) and Cornfield (1971).

[10] Even though we have hypothesized τ_j for all trials, we cannot calculate hypothesized $\bar{y}_j(E)$ and $\bar{y}_j(C)$ for the unsampled trials, and thus the probabilistic statement is conditional on the observed trials.

additional variables(s) still depends on the appropriateness of the underlying linear model.

Hence, given random sampling of trials, the ability to generalize results to other trials seems relatively straightforward probabilistically. However, most experiments are designed to be generalized to *future* trials; we never have a random sample of trials from the future but at best a random sample from the present; in fact, experiments are usually conducted in constrained, atypical environments and within a restricted period of time. Thus, in order to generalize the results of any experiment to future trials of interest, we minimally must believe that there is a similarity of effects across time and more often must believe that the trials in the study are "representative" of the population of trials. This step of faith may be called making an assumption of "subjective random sampling" in order to assert such properties as (*a*) $\bar{y}_d$ (or $\bar{y}_d$ adjusted) tends to estimate the typical causal effect T and (*b*) the plausibility of hypothesized $\bar{\tau}_j$, $j = 1, \cdots, M$, is given by the usual conditional significance level.

Even though the trials in an experiment are often not very representative of the trials of interest, investigators do make and must be willing to make this assumption of subjective random sampling in order to believe their results are useful. When investigators carefully describe their sample of trials and the ways in which they may differ from those in the target population, this tacit assumption of subjective random sampling seems perfectly reasonable. If there is an important variable that differs between the sample of trials and the population of trials, an attempt to adjust the estimate based on the same kinds of models discussed previously is quite appropriate.[11]

Presenting the Results of an Nonrandomized Study as Being of General Interest

The same two issues previously discussed as arising when presenting the results of an experiment also arise when presenting the results of a nonrandomized study as being relevant. However, the first issue, the effect of variables not explicitly controlled, is usually more serious in nonrandomized than in randomized studies, while the second, the applicability of the results to a population of interest, is often more serious in randomized than in nonrandomized studies.

Effect of Variables Not Explicitly Controlled

In order to believe that $\bar{y}_d$ in a nonrandomized study is a good estimate of τ, the typical causal effect for the $2N$ trials in the study, we must believe that there are no extraneous variables that affect Y and systematically differ in the E and C groups; but we have to believe this even in a randomized experiment. The primary difference is that without randomization there is often a strong suspicion that there are such variables, while with randomization such suspicions are generally not as strong.

Consider a carefully controlled nonrandomized study—a study in which there are no obviously important prior variables that systematically differ in the E trials and the C trials. In such a study, there is a real sense in which a claim of "subjective randomization" can be made. For example, if the study was composed of carefully matched pairs of trials, there might be a very defensible belief that within each matched pair each unit was equally likely to receive E as C in the sense that if you were shown the units without being told which received E, only half the time would you correctly guess which received E.[12] Under this assumption of subjective randomization, the usual estimates and significance levels can be used as if the study had been randomized; this procedure is analogous to assuming subjective random sampling in order to make inferences about a target population. Until an obviously important variable is found that systematically differs in the E and C trials, the belief in subjective randomization is well founded.

Now consider a nonrandomized study in

[11] See Cochran (1963) on regression and ratio adjustments. These are appropriate whether the sample is actually random or not.

[12] Perhaps this is all that is meant by "randomization" to some Bayesians under any circumstance (see Savage, 1954, p. 66).

which an obviously important prior variable is found that systematically differs in the E and C trials. We must adjust the estimate $\bar{y}_d$ and the associated significance levels just as we would if the study were in fact a properly randomized experiment. An obvious way to adjust for such variables is to assume subjective randomization (i.e., the study was randomized and the observed difference on prior variables occurred "by chance"), and use the methods discussed in the previous section "Considering Additional Variables" appropriate for an experiment (i.e, gain scores, adjustment by a known function, covariance adjustment).

The main problem with this approach is that having found an important prior variable that systematically differs in the E and C trials, we might suspect that there are other such variables, while if the study were randomized we might not be as suspicious of finding these prior variables. Additionally, the various methods of adjustment that yield unbiased estimates given randomization have varying biases under different models without randomization. Even though an unbiased estimate is not (as we have seen) the total answer to estimating τ, it is more desirable than a badly biased estimate. Recent work on methods of reducing bias in nonrandomized studies is summarized in Cochran and Rubin (1974). Much work remains to be done, especially for many prior variables and nonlinear relations between these and Y.

In sum, with respect to variables not explicitly controlled, a randomized study leaves the investigator in a more comfortable position than does a nonrandomized study. Nevertheless, the following points remain true for both: (*a*) Any adjustment is somewhat dependent upon the appropriateness of the underlying model—if the model is appropriate the confounding effect of the prior variables is reduced or eliminated, while if the model is inappropriate a confounding effect remains. (*b*) We can never know that all causally relevant prior variables that systematically differ in the E and C trials have been controlled. (*c*) We must be prepared to ignore irrelevant prior variables even if they systematically differ in E and C trials, or else we can obtain any estimate desired by eventually finding the "right" irrelevant prior variables.

Generalizing Results to Other Trials

For almost any study to be of interest, the results must be generalizable to a population of trials. Typically, nonrandomized studies have more representative trials than experiments since these are often conducted in constrained environments. Thus, if the choice is between a nonrandomized study whose $2N$ trials consisted of N representative E trials closely matched to N representative C trials and an experiment whose $2N$ trials were highly atypical, it is not clear which we should prefer; in practice there may be a trade-off between the reasonableness of the assumptions of subjective random sampling and subjective randomization (e.g., consider a carefully matched nonrandomized evaluation of existing compensatory reading programs and an experiment having these compensatory reading programs randomly assigned to inmates at a penitentiary).

In a sense, all studies lie on a continuum from irrelevant to relevant with respect to answering a question. A poorly controlled nonrandomized study conducted on atypical trials is barely relevant, but a small randomized study with much missing data conducted on the same atypical trials is not much better. Similarly, a very well-controlled experiment conducted on a representative sample of trials is very relevant, and a very well-controlled nonrandomized study (e.g., E and C trials matched on all causally important variables, several control groups each with a potentially different bias) conducted on a representative sample of trials is almost as good. Typically, real-world studies fall somewhere in the middle of this continuum with nonrandomized studies having more representative trials than experiments but less control over prior variables.

Summary

The basic position of this paper can be summarized as follows: estimating the typical causal effect of one treatment versus another is a difficult task unless we understand the actual process well enough to (*a*) assign most of the variability in Y to specific

causes and (*b*) ignore associated but causally irrelevant variables. Short of such understanding, random sampling and randomization help in that all sensible estimates tend to estimate the correct quantity, but these procedures can never completely assure us that we are obtaining a good estimate of the treatment effect.[13]

Almost never do we have a random sample from the target population of trials, and thus we must generally rely on the belief in subjective random sampling, that is, there is no important variable that differs in the sample and the target population. Similarly, often the only data available are observational and we must rely on the belief in subjective randomization, that is, there is no important variable that differs in the E trials and C trials. With or without random sampling or randomization, if an important prior variable is found that systematically differs in E and C trials or in the sample and target population, we are faced with either adjusting for it or not putting much faith in our estimate. However, we cannot adjust for any variable presented, because if we do, any estimate can be obtained.

In both randomized and nonrandomized studies, the investigator should think hard about variables besides the treatment that may causally affect Y and plan in advance how to control for the important ones—either by matching or adjustment or both. When presenting the results to the reader, it is important to indicate the extent to which the assumptions of subjective randomization and subjective random sampling can be believed and what methods of control have been employed.[14] If a nonrandomized study is carefully controlled, the investigator can reach conclusions similar to those he would reach in a similar experiment.[15] In fact, if the effect of the E versus C treatment is large enough, he will be able to detect it in small, nonrepresentative samples and poorly controlled studies.

Basic problems in social science research are that causal models are not yet well formulated, there are many possible treatments, and in many cases the differential effects of treatments appear to be quite small. Given this situation, it seems reasonable to (*a*) search for treatments with large effects using well-controlled nonrandomized studies when experiments are impractical and (*b*) rely on further experimental study for more refined estimates of the effects of those treatments that appear to be important. The practical alternative to using nonrandomized studies in this way is evaluating many treatments by introspection rather than by data analysis.

[13] Even assuming a good estimate of the causal effect of E versus C, there remains the problem of determining which aspects of the treatments are responsible for the effect. Consider, for example, "expectancy" effects in education (Rosenthal, 1971) and the associated problems of deciding the relative causal effects of the content of programs and the implementation of programs.

[14] Recent advice on the design and analysis of observational studies is given by Cochran in Bancroft (1972).

[15] See, for example, the large scale evaluation of Salk vaccine described by Meier in Tanur et al (1972).

REFERENCES

Bancroft, T. A. *Statisical papers in honor of George W. Snedecor*. Ames: Iowa State University Press, 1972.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *The disadvantaged child*. Vol. 3. *Compensatory education: A national debate*. New York: Brunner/Mazel, 1970.

Cochran, W. G. *Sampling techniques*. New York: Wiley, 1963.

Cochran, W. G., & Cox, G. M. *Experimental designs*. (6th ed.) New York: Wiley, 1957.

Cochran, W. G., & Rubin, D. B. Controlling bias in observational studies: A review. Mahalanobis Memorial Volume *Sankhyā -A*, 1974, 1–30.

Cornfield, J. The University Group Diabetes Program. *Journal of the American Medical Association*, 1971, **217**, 1676–1687.

Davies, O. L. *Design and analysis of industrial experiments*. London, England: Oliver & Boyd, 1954.

Fisher, R. A. *Statistical methods for research workers*. (1st ed.) New York: Hafner, 1925.

Kempthorne, O. *The design and analysis of experiments*. New York: Wiley, 1952.

Lehmann, E. L. *Testing statistical hypotheses*. New York: Wiley, 1959.

Rosenthal, R. Teacher expectation and pupil learning. In R. D. Strom (Ed.), *Teachers and the*

learning process. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Savage, L. J. *The foundations of statistics.* New York: Wiley, 1954.

Scheffé, H. *The analysis of variance.* New York: Wiley, 1959.

Schor, S. The University Group Diabetes Program. *Journal of the American Medical Association,* 1971, **217,** 1671–1675.

Tanur, J. M., Mosteller, F., Kruskal, W. H., Link, R. F. Pieters, R. S., & Rising, G. R. *Statistics: A guide to the unknown.* San Francisco, Calif.: Holden Day, 1972.

(Received July 30, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 702-709

RELATIONSHIP OF THREE COGNITIVE STYLE DIMENSIONS TO ELEMENTARY READING ABILITIES[1]

DOUGLAS R. DENNEY[2]

University of Kansas

Good and poor readers from Grades 2 to 5 were compared on three cognitive style dimensions—conceptual style preferences, cognitive tempos, and attentional styles—which were assessed with the Conceptual Styles Test, Matching Familiar Figures Test, and Fruit Distraction Test. Attentional style measures distinguished poor and good readers better than the other cognitive style measures. However, another set of scores aligned with but not identical to the attentional style measures were even more highly related to reading. These scores reflected children's skills at sequentially transposing information from visual to verbal channels. The results are considered in light of the maturational lag hyopthesis; alternative accounts are proposed for the changing patterns of deficits between younger and older poor readers.

Elementary grade children show dramatic developmental advances along several cognitive style dimensions (Goodenough & Eagle, 1963; Kagan, 1966; Santostefano & Paley, 1964). The aim of the present study was to evaluate the relationship of three such dimensions—conceptual style preferences, cognitive tempos, and attentional styles—to reading abilities.

Conceptual style preference (Kagan, 1966) refers to the type of conceptual relationships among objects typically formed by the child. Three such preferences that have been identified are analytic, relational, and inferential conceptualization. An analytic concept is one which is formed on the basis of an objective element common to all members included in the group (e.g., they all have sharp points). A relational concept is based upon a functional relationship established among the members (e.g., the match is used to light the pipe). An inferential concept is based upon the members' belonging to some superordinate class identified by a lexical term (e.g., they are both tools).

Increases in preferences for analytic and inferential concepts and decreases in relational concepts with increasing age are reported among elementary children (Kagan, 1966). Kagan has cited increases in the child's tendency to analyze stimuli into their component parts as partly responsible for the increase in analytic conceptual style. Others have seen the shift from a focus upon "things that go together" to "things that are alike" as reflective of a more basic cognitive attainment by the child (Denney, 1971; Denney & Lennon, 1972). In either case, analytic and inferential conceptual functioning may be related positively to reading ability since visual analysis of words and a focus upon words and word parts, which are alike and thus perhaps sound alike, would both seem to aid reading acquisition, particularly within phonemic instructional programs.

The cognitive tempo dimension involves the child's tendency to inhibit initial responses and to reflect upon the accuracy of

[1] This study was supported by a National Institute of Mental Health, Small Grant (MH 23656-01). The author acknowledges the assistance of Lee Smith, principal, and Jean Harmon, special reading instructor, Tongonoxie Public School, Tongonoxie, Kansas. The author also thanks Rosemary Wright, Brenda Morris, and Mark Devaney, who served as research assistants.

[2] Requests for reprints should be sent to Douglas R. Denney, Department of Psychology, University of Kansas, Lawrence, Kansas 66045.

his answer rather than responding impulsively, in conditions of high response uncertainty. Developmental studies reveal a trend from impulsive to reflective cognitive tempos (Kagan, 1966).

The identification of words and the reading of sentences by children just acquiring the ability to read represent situations of high response uncertainty within which cognitive tempos might be expected to have relevance. Using a sample of first-grade children, Kagan (1965) found positive correlations between indices of reflectivity and word recognition assessed both at the same time and six months after the cognitive tempo measures. Negative correlations were reported between some of the same indices of reflectivity and the errors committed during the oral reading of paragraphs a full year after the cognitive tempo measures were taken. Kagan recommended that measures of cognitive tempo be incorporated in diagnostic batteries for the early identification of poor readers.

Attentional style refers to the child's ability to deploy his attention selectively, thereby avoiding distraction from intrusive and irrelevant stimulus information. Two such styles are constriction and flexibility, children with the flexible attentional style being less distracted by irrelevant stimuli. Santostefano and Paley (1964) showed developmental advances along this dimension from constricted to flexible attentional styles.

Effective reading would seem to necessitate the accurate deployment of attention and the exclusion of irrelevant and distracting information either contained within the relevant stimuli (e.g., silent letters) or in close proximity to them (e.g., pictures and other words on the page). Santostefano, Rutledge, and Randall (1965) found constriction–flexibility to be the only one of three cognitive style dimensions (the others being focusing–scanning and leveling–sharpening) to distinguish between poor and good readers.

The present study constitutes an extension of both Kagan's (1965) and Santostefano et al.'s (1965) studies in that (*a*) a new cognitive style dimension, conceptual style preferences, is added; (*b*) unlike Santostefano et al.'s study, many specific reading skills are assessed separately and the relationship of each to the cognitive style dimensions is evaluated; and (*c*) unlike Kagan's study, the relevance of each cognitive style dimension is investigated over the course of the several elementary grades during which reading ability is gradually being acquired.

Reading deficiences have been thought to reflect maturational lags in the Central Nervous System, with poor readers manifesting cognitive–perceptual behaviors similar to those of younger normal children (Satz, Rardin, & Ross, 1971). It follows that the deficits distinguishing older poor readers from the age-mates should be different from those distinguishing younger poor readers from their age-mates. Since dramatic attainments in perceptual–motor and attentional skills precede major shifts in language and conceptual–symbolic skills, deficits in the former skills should be more prevalent among younger poor readers and deficits in the latter skills more prevalent among older poor readers. A number of studies comparing poor and normal readers at more than one age level lend support to the above hypothesis (Braun, 1963; Satz et al., 1971).

The present study employed children ranging from the second through the fifth grades and included measures reflecting a wide variety of cognitive–perceptual behaviors. Thus, it was possible to evaluate the results of the study in terms of the shift from perceptual–motor and attentional deficits to language, conceptual, and intellective deficits posited by the maturational lag hypothesis.

Method

Subjects

The subjects were white, middle-class children in the second through the fifth grades of a small, rural elementary school. Ten children from each grade were designated poor readers on the basis of their reading teacher's judgment; all 40 of these children were receiving weekly special reading instruction. The majority of these children were boys; however, there was 1 girl from the second grade, 2 from both the third and fifth grades, and

3 from the fourth grade. Ten average or slightly above-average readers from each grade were also selected, again based upon the reading teacher's judgment. Highly advanced readers were excluded from the study. Within each grade level, poor and normal readers were matched for sex. Beyond these constraints, the selection of normal readers to compare with the poor readers was done randomly from a larger list of average or above-average readers.

Tests

The 80 subjects were individually administered Form C of the Gilmore Oral Reading Test (Gilmore & Gilmore, 1968) and four subtests from the Gates-McKillop Reading Diagnostic Tests (Gates & McKillop, 1962). The Gilmore Oral Reading Test consisted of a series of paragraphs, graduated in difficulty, which the child read aloud. The examiner tallied the number of reading errors committed on each passage, the total being converted into an accuracy raw score. In addition, the child was asked three questions over the material in each passage. The total number of questions answered correctly for all passages became the child's comprehension raw score. Finally, the child's reading rate in words per minute was also computed.

The four subtests from the Gates-McKillop were the following:

Word Pronunciation—Untimed Presentation (Whole Words). This test consisted of a list of 80 words ordered in terms of difficulty. Two attempts to read each word were allowed, with points awarded for correct pronunciations during the first or second trials.

Recognizing and Blending Common Word Parts (Word Parts). Twenty-three nonsense words, each composed of two common phonetic elements (e.g., SKED: SK, and ED), were presented. Four points were awarded for correctly reading the whole word on the first trial; if this was not done, one point each was awarded for correctly reading the first and second elements and for correcly blending them in a final pronunciation of the word.

Recognizing the Visual Forms or Word Equivalents of Sounds (Nonsense Words). After hearing each of 20 nonsense words, the child circled 1 of 4 written words matching the sound he had just heard.

Auditory Blending. The component sounds for 15 common words were presented at one-fourth-of-a-second intervals. The child was awarded one point for each word correctly pronounced within two attempts.

Subjects were also individually administered the following cognitive style tests.

Conceptual Styles Test—Form A (CST-A). This test, adapted from Kagan's Conceptual Styles Test (Kagan, 1966), provided a measure of children's preferences for analytic, relational, and inferential concepts. The CST-A was composed of 15 items, each containing three pictures from which the subject was instructed to choose two "that are alike or go together in some way." Items were designed so that one possible pair represented an analytic concept and another represented a relational concept. The child's pointing response and accompanying explanation were categorized as analytic, relational, or inferential (Denney, 1971).

Matching Familiar Figures Test (MFFT). This test, designed by Kagan to assess the cognitive tempos of reflection and impulsivity, consisted of 14 match-to-standard items, including 2 practice items. Each item consisted of a standard and 6 alternative stimuli exposed simultaneouslv. Latency to the first response and total number of errors were summed across all 12 test items. Both latency and error scores were used as indices of cognitive tempo (Kagan, 1966).

Fruit Distraction Test (FDT). The third test was adapted from one designed by Santostefano and Paley (1964) to assess constricted and flexible attentional styles. The test consisted of three cards, each containing 50 pictures of bananas, cherries, grapes, and carrots ordered randomly on the card in 10 rows of 5. On Card 1, the fruits and vegetables were colored appropriately (yellow, red, purple, and orange, respectively). Card 2 was identical to Card 1 with the addition of a number of achromatic drawings of common objects interspersed among the relevant stimuli. Card 3 was identical to Card 1 except that the fruits and vegetables were all colored inappropriately. On Cards 1 and 2, the child was required to name the colors of each fruit and vegetable, and on Card 3, he was required to name the color that each fruit and vegetable should be, proceeding in all cases from left to right down each card. "Reading" times and numbers of corrected and uncorrected errors were recorded for each card. Following the reading of Card 2, the subject was asked to recall as many of the achromatic drawings as he could. Differences in reading times and errors between Card 2 and Card 1, and between Card 3 and Card 1, and the number of achromatic drawings recalled on Card 2 constituted five indices of attentional style.

Finally, Form A of the Peabody Picture Vocabulary Test (PPVT) was administered to each child. The PPVT is a test of receptive vocabulary correlating highly with other global measures of intelligence (Dunn, 1965).

Three sessions were required to administer all the measures to each child. The first two sessions were conducted during the middle of the school year, with no more than two weeks separating the two sessions for any child. During the first session, a female examiner administered the Gilmore; during the second, a male examiner administered the CST-A, MFFT, FDT, and Gates-McKillop tests. The PPVT was administered by a female examiner during a third session held at the end of the school year.

Design

The study conformed to a 4 × 2 (Grade × Reading Level) factorial design with 10 subjects in

each cell. To control for intellectual differences among the subjects in the evaluation of the reading and cognitive style data, an analysis of covariance was employed, using PPVT IQ scores as the covariate. Similarly, in correlating reading skill and cognitive style measures, intellective differences were removed through partial correlation.

Results

Analysis of the Reading Measures

A 4 × 2 (Grade × Reading Level) analysis of covariance was performed on each of the seven reading scores derived from the Gilmore Oral Reading Test and the Gates-McKillop Reading Diagnostic Tests. Table 1 presents the adjusted group means for both poor and good readers within each of the grades on the seven reading scores.

Highly significant differences between poor and good readers were discovered on all seven reading measures with *F*s significant at the .001 level for all measures except auditory blending: (Gilmore) accuracy ($F = 124.34$, $df = 1/71$, comprehension ($F = 16.87$, $df = 1/71$), rate ($F = 183.90$, $df = 1/71$); (Gates-McKillop) whole words ($F = 26.62$, $df = 1/71$), word parts ($F = 44.85$, $df = 1/71$), nonsense words ($F = 52.23$, $df = 1/71$), auditory blending ($F = 2.85$, $df = 1/71$, $p < .025$). In addition, there were significant Grade × Reading Level interactions on five of these measures: accuracy ($F = 4.33$, $df = 3/71$, $p < .01$); rate ($F = 2.86$, $df = 3/71$, $p < .05$); word parts ($F = 3.55$, $df = 3/71$, $p < .025$); nonsense words ($F = 5.64$, $df = 3/71$, $p < .005$); auditory blending ($F = 2.85$, $df = 3/71$, $p < .05$). Separate comparisons of good and poor readers within each grade level were performed for these five measures to identify instances in which the differences between poor and good readers increased or decreased consistently across the grades. Such trends were found in the case of the accuracy and rate data. Although poor readers performed significantly lower than good readers in both accuracy and rate at each grade level (all *p*s < .001), the differences tended to be greater within the latter two grades.

Analysis of the Cognitive Style Measures

As with the reading measures, the cognitive style measures were analyzed using a 4 × 2 analysis of covariance, with PPVT IQ scores as the covariate. Table 2 presents the adjusted group means for poor and good readers on 11 variables derived from the cognitive style measures.

Analyses of the frequencies of analytic, relational, and inferential concepts on the CST-A revealed no significant results for reading level or Grade × Reading Level interaction. However, the interactions for both the analytic and the relational scores approached significance (analytic: $F = 2.57$, $df = 3/71$, $p < .10$; relational: $F = 2.39$, $df = 3/71$, $p < .10$), prompting specific comparisons for these scores. In the fifth grade, poor readers gave significantly fewer analytic responses ($F = 4.77$, $df = 1/17$, $p < .05$) and significantly more relational

TABLE 1

Adjusted Group Means for the Gilmore Oral Reading Test and Gates-McKillop Reading Diagnostic Test Measures

Reading measure	2nd		3rd		4th		5th	
	Poor	Good	Poor	Good	Poor	Good	Poor	Good
Gilmore								
Accuracy	5.3	26.7	16.6	38.6	15.4	50.1	20.4	57.5
Comprehension	10.4	19.9	19.4	21.5	20.4	28.1	21.4	32.3
Rate	23.4	105.8	46.6	109.7	43.9	122.9	53.5	154.9
Gates-McKillop								
Whole Words	34.7	58.8	44.6	72.4	63.6	77.4	44.3	80.1
Word Parts	4.0	72.8	54.4	82.4	51.2	90.0	47.2	93.2
Nonsense Words	9.1	17.4	13.6	17.5	15.4	18.2	13.4	18.4
Auditory Blending	12.2	14.0	12.0	13.8	9.5	14.7	14.2	14.3

TABLE 2

ADJUSTED GROUP MEANS FOR ALL SCORES DERIVED FROM THE COGNITIVE STYLE MEASURES

Cognitive style measure	2nd		3rd		4th		5th	
	Poor	Good	Poor	Good	Poor	Good	Poor	Good
CST-H								
Analytic	9.6	7.6	6.8	8.3	6.9	4.2	4.2	9.2
Relational	2.7	4.9	6.5	5.2	6.1	8.9	9.8	4.9
Inferential	2.5	4.2	1.0	1.6	1.7	1.8	0.8	1.3
MFFT								
Latency	78.4	118.0	97.0	94.7	103.0	91.0	96.2	102.6
Error	17.2	16.5	17.4	16.2	16.6	15.7	15.2	12.8
FDT								
Rd Time 1	131.5	92.7	110.2	104.9	115.7	73.1	97.7	68.5
Error 1	0.7	1.4	0.4	0.4	2.5	0.4	0.2	0.9
Rd Time 2	151.7	90.0	110.4	99.0	109.8	64.3	97.4	65.7
Error 2	2.7	0.7	1.1	0.5	2.4	0.4	0.8	0.2
Rd Time 3	200.6	153.6	170.2	179.8	217.8	119.9	129.7	121.3
Error 3	1.8	0.4	0.6	1.0	3.7	0.5	1.5	1.2

Note. Abbreviations: CST-A = Conceptual Styles Test - Form A; MFFT = Matching Familiar Figures Test; FDT = Fruit Distraction Test; Rd = Reading.

responses ($F = 4.51$, $df = 1/17$, $p < .05$) than good readers. There were no differences in the earlier grades.

Analyses of the latency and error scores on the MFFT revealed no significant results.

Analyses of the difference scores between Card 2 and Card 1 on the FDT revealed a significant reading level main effect for the error difference scores ($F = 5.24$, $df = 1/71$, $p < .05$) and a significant Grade × Reading Level interaction for the reading time difference scores ($F = 3.28$, $df = 3/71$, $p < .05$). There was also a significant Grade × Reading Level interaction for the reading time difference scores between Card 3 and Card 1 ($F = 3.18$, $df = 3/71$, $p < .05$). Separate comparisons performed on these interactions failed to show any consistent trends in the degree of disparity between good and poor readers across the four grade levels. No significant findings resulted from the analysis of the number of achromatic drawings recalled on Card 3.

Sizable differences between poor and good readers were apparent in both reading times and errors on each individual card comprising the FDT, and analyses of covariance were therefore performed on these data. Significant reading level main effects were found for Card 1 reading times ($F = 16.34$, $df = 1/71$, $p < .001$), Card 2 reading times ($F = 31.55$, $df = 1/71$, $p < .001$), and errors ($F = 6.30$, $df = 1/71$, $p < .025$). Significant Grade × Reading Level interactions were found on Card 2 reading times ($F = 4.71$, $df = 3/71$, $p < .01$), Card 3 reading times ($F = 3.23$, $df = 3/71$, $p < .05$), and errors ($F = 3.25$, $df = 3/71$, $p < .05$). Separate comparisons performed on these interactions failed to reveal any trends in the degree of disparity between good and poor readers across the grade levels.

Partial correlations were computed between the cognitive style measures and the reading measures, for the total sample of 80 children as well as for the 40 second- and third-grade children and the 40 fourth- and fifth-grade children. There were no significant correlations between the CST-A scores or the MFFT latency scores and the reading measures. MFFT error scores correlated significantly with accuracy ($r = -.238$), comprehension ($r = -.291$), and rate ($r = -.247$) on the Gilmore, for the total sample but not for the separate samples of older and younger children.

Substantially larger correlations were found between FDT scores and the reading measures.[3] Time difference and error

[3] A table of correlations between the Fruit Distraction Test (FDT) and reading measures is available from the author upon request.

difference scores between Card 2 and Card 1 were more highly correlated with the reading measures (average $r = -.202$) than were the corresponding difference scores between Card 3 and Card 1 (average $r = -.161$). Similarly, reading times and errors on Card 2 were more highly related to the reading measures (average $r = -.298$) than the corresponding scores on Card 3 (average $r = -.202$). Reading measures tended to correlate more highly with reading time and error scores on the individual cards than with time difference and error difference scores computed between cards. In particular, the reading time scores on Card 1 and Card 2 and the error scores on Card 2 were related to reading measures for the total sample. The reading time scores on Card 1 correlated significantly with every reading measure except whole words (*rs* ranging from $-.145$ to $-.465$). The reading time scores and Card 2 correlated significantly with every reading measure (*rs* ranging from $-.224$ to $-.549$). The error scores on Card 2 correlated significantly with every reading measure except auditory blending and word parts (*rs* ranging from $-.063$ to $-.420$). For both the individual card scores and the difference scores on the FDT, correlations were consistently higher for younger children (average $r = .235$) than for older children (average $r = .174$).

Discussion

The analyses of the reading measures confirm substantial differences between the poor and good readers in all of their reading skills. The fact that the poor readers were receiving special reading instruction and were thus identified as poor readers to themselves and their peers is a potential source of confounding in this study. This difference between groups may have been amplified by the stigma attached to the poor readers or may have been moderated by the special training they were receiving; however, this confounding was unavoidable given the available subjects.

Conceptual style preferences generally failed to distinguish between poor and good readers, the only exception being for the fifth-grade children. In this case, poor readers tended to employ the developmentally less advanced relational concepts rather than analytic concepts.

Cognitive tempo data failed in all instances to distinguish between good and poor readers. Kagan's (1965) findings were upheld only in the case of three significant but moderate correlations between MFFT errors and Gilmore Oral Reading Test scores. Eleven other correlations between reading and cognitive tempo measures failed to attain significance, raising serious doubt about the importance of this dimension to reading ability.

In contrast to the conceptual style and cognitive tempo dimensions, substantial differences between good and poor readers were found along the attentional style dimension. However, the results contradict those of Santostefano et al. (1965). In the present study, difference scores between Card 2 and Card 1 successfully distinguished good and poor readers, while in Santostefano's study, difference scores between Card 3 and Card 1 were found to relate to reading. There appears to be no way to resolve these different outcomes. Together, the studies point up the necessity of assessing attention deployment in the face of distracting information both embedded within the relevant stimulus (Card 3) and in close proximity to it (Card 2).

In terms of developmental attainments along each cognitive style dimension, poor readers share with their normal peers the same orientations toward the similarities common to visual stimuli and the same abilities to analyze such stimuli into their component parts, skills which are clearly necessary in early reading acquisition. Poor readers also demonstrate the same concerns for the adequacy of their responses and the same tendencies to delay initial responses in order to reflect over their answers; without these attainments in reflection, initial attempts at reading would amount to nothing more than random guessing with too little time allotted between stimulus and response for the child to apply any of what he may have learned from his reading program. The present study indicates that poor readers' difficulties lie not in the total amount of time they attended to particular problems but in the proportion of that time

spent productively examining the relevant stimuli in their visual field. Poor readers' failures to focus upon the relevant stimuli not only hinder them in the act of reading itself (by distracting them from the word or phrase to be read) but also in the process of learning how to read. Staats, Brewer, and Gross (1970) have shown that differences in the learning curves for good and poor readers being taught the letters of the alphabet through operant conditioning exist in the early trials; during such trials, the direction of attention to the learning task at hand is a most crucial factor affecting the shape of the learning curves.

Reading time and error scores on the individual cards of the FDT differentiated between good and poor readers more effectively than any of the three cognitive style dimensions. The average correlation between all difference scores and their respective individual card scores was .485, indicating the cognitive–perceptual behavior assessed by the individual card scores to be related but not identical to the dimension of constriction–flexibility. Instead, the individual card scores seem to reflect the child's ability to process visually inputed information and to encode his response in a verbal channel. This sequential, visual-to-verbal transposition shares many similarities with the act of reading, and failures in intermodal integration between information-processing channels have frequently been implicated in poor reading (Birch & Belmont, 1965; Bryden, 1972).

There is some evidence of a shift from perceptual–motor and attentional deficits among younger poor readers to language and conceptual deficits among older poor readers, a finding consistent with the maturational lag hypothesis. Performance of the FDT was more highly related to reading scores for younger children, while scores on the CST-A and the PPVT were more highly related to reading scores for older children. However, the differences in the correlations for older and younger children are not large and the failure to find Grade × Reading Level interactions consistent with trends predicted from the maturational lag hypothesis must detract further from its support in the present study.

Even if the evidence in this study was more supportive of the predicted shift in deficits from younger to older poor readers, such a shift might be explained by learning rather than by a genetically determined and organically tied maturational lag. Failures to learn earlier skills may impede acquisition of later skills. To date there is no compelling evidence that the attentional and visual–verbal integrative skills related to reading in the present study are genetically determined.

There is yet another way to account for differences in the nature of the deficits shown by poor readers of various ages. Evidence of these differences is typically derived from cross-sectional studies of groups of poor readers and there may well be differences in the composition of these groups from one grade to the next. A fair number of poor readers respond quickly to early special reading instruction during the first two of three grades and perform at average reading levels thereafter. Others continue in special reading, profiting relatively little from these programs and thus falling further behind with each successive year. Groups of poor readers may therefore be composed of two types of children: (*a*) those who have relatively minor perceptual–motor and attentional deficits which respond to specialized instructional techniques and those with more severe cognitive–intellective and language dysfunctions which fail to respond to specialized reading and which indeed become aggravated as the child continues without the expanded experience made available through the facility of reading. In subsequent grade levels, the proportion of the former type of poor reader may decrease, while the deficits of the latter type of poor reader become more severe. Longitudinal studies are needed in further investigations of the maturation lag hypothesis. In addition, investigators might begin examining the ability to profit from reading instruction as a critical dimension distinguishing among poor readers.

REFERENCES

Birch, H. G., & Belmont, L. Auditory-visual integration, intelligence, and reading ability in school children. *Perceptual and Motor Skills*, 1965, **20**, 295–305.

Braun, J. S. Relation between concept formation ability and reading achievement at three developmental levels. *Child Development*, 1963, **34**, 675–682.

Bryden, M. P. Auditory-visual and sequential-spatial matching in relation to reading ability. *Child Development*, 1972, **43**, 824–832.

Denney, D. R. The assessment of differences in conceptual style. *Child Study Journal*, 1971, **1**, 142–155.

Denney, N. W., & Lennon, M. Classification: A comparison of middle and old age. *Developmental Psychology*, 1972, **7**, 210–213.

Dunn, L. M. *Manual for the Peabody Picture Vocabulary Test*. Circle Pines, Minn.: American Guidance Service, 1965.

Gates, A. I., & McKillop, A. S. *Manual for the Gates-McKillop Reading Diagnostic Tests*. New York: Columbia University, Teachers College, Bureau of Publications, 1962.

Gilmore, J. V., & Gilmore, E. C. *Manual for the Gilmore Oral Reading Test*. New York: Harcourt, Brace, and World, 1968.

Goodenough, D. R., & Eagle, C. J. A modification of the Embedded Figures Test for use with young children. *Journal of Genetic Psychology*, 1963, **103**, 67–74.

Kagan, J. Reflection-impulsivity and reading ability in primary grade children. *Child Development*, 1965, **36**, 609–628.

Kagan, J. Developmental studies on reflection and analysis. In A. Kidd & J. Rivoire (Eds.), *Perceptual and conceptual development in children*. New York: International Universities Press, 1966.

Santostefano, S., & Paley, E. Development of cognitive controls in children. *Child Development*, 1964, **35**, 939–949.

Santostefano, S., Rutledge, L., & Randall, D. Cognitive styles and reading disabiltiy. *Psychology in the Schools*, 1965, **2**, 57–62.

Satz, P., Rardin, D., & Ross, J. An evaluation of a theory of specific developmental dyslexia. *Child Development*, 1971, **42**, 2009–2021.

Staats, A. W., Brewer, B. A., & Gross, M. C. Learning and cognitive development: Representative samples, cumulative-hierarchical learning, and experimental-longitudinal methods. *Monographs of the Society for Research in Child Development*, 1970, **35** (8, Serial No. 141).

(Received September 10, 1973)

Manuscripts Accepted for Publication in the *Journal of Educational Psychology*

Locus of Control and Aptitude Test Scores as Predictors of Academic Achievement. Walter R. Nord (Graduate School of Business Administration, Washington University, St. Louis, Missouri 63130), Francis Connelly, and George Daignault.

Locus of Control, Test Anxiety, and Student Performance in a Personalized Instruction Course. George J. Allen (Department of Psychology, University of Connecticut, Storrs, Connecticut 06268), Laura Giat, and Roger J. Cherney.

Interaction of Learner Aptitudes with Question Pacing in Learning from Prose. Mary Lou Koran (Foundations of Education Department, Institute for Development of Human Resources, 513 Well Hall, College of Education, University of Florida, Gainesville, Florida 32611) and John J. Koran, Jr.

Serial Memory Span Thresholds of Normal and Mentally Retarded Children. Alfred A. Baumeister (The John F. Kennedy Center for Research on Education and Human Development, George Peabody College for Teachers, Nashville, Tennessee 87203).

Effect of Two Free Time Reinforcement Procedures on Academic Performance in a Class of Behavior Problem Children. David Marholin 11 (106 West Belle Street, P. O. Box 187, Oreana, Illinois 62554), Elizabeth T. McInnis, and Tom B. Heads.

Retention of Pictorial and Verbal Content of a Text with Illustrations. Joan Peeck (Psycholisch Laboratorium der Rijksuniversiteit, Varkenmarkt 2, Utrecht, The Netherlands).

Reading Level Grade Adjustment and Auditory Supplementation as Techniques for Increasing Textual Comprehensibility. Arthur I. Siegel (Applied Psychological Services, Science Center, Wagne, Pennsylvania) and Martin R. Lautman.

Influence of Cultural Background on Hierarchical Learning. Russell D. Linke (Faculty of Education, Monash University, Clayton, Victoria 3168, Australia).

Outerdirectedness in Children of Three ages as a Function of Experimentally Induced Success and Failure. Donald L. MacMillan (Department of Education, University of California, Riverside, California 92502) and Deborah L. Wright.

Information-Seeking Following the Confirmation or Contradiction of Beliefs. Charles B. Schultz (Trinity College, Hartford, Connecticut 06106).

Effect of Hypothesis/Test Training on Reading Skill. S. Jay Samuels, (Center for Research in Human Learning, University of Minneapolis, Elliott Hall, Minneapolis, Minnesota 55455), Patricia Dahl, and Teara Archwamety.

Training Letter Discrimination by Presentation of High Confusion versus Low Confusion Alternatives. Rosemary O. Nelson (Department of Psychology, University of North Carolina, Greensboro, North Carolina 27412) and Kenneth S. Wein.

A Multitrait-Multimethod Analysis of Conceptual Tempo. Vernon C. Hall (Syracuse University, 331 Huntington Hall, Syracuse, New York 13210) and William J. C. Russell.

(*Continued on p. 792*)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 710-717

EFFECTS OF COGNITIVE INTERVENTION ON LATENT ABILITY MEASURED FROM ANALOGY ITEMS[1]

SUSAN E. WHITELY[2] AND RENÉ V. DAWIS

University of Minnesota

Cognitive intervention may potentially be used to investigate both the cognitive processes underlying test performance and the possible incorporation of learning experiences into ability estimates. Inner-city high school students were randomly assigned to one of five experimental conditions or to the control group and were administered a test–intervention–retest sequence using verbal analogy items. It was found that although latent ability estimates may be significantly increased by a short intervention, modifiability depends on the type of intervention. Furthermore, it was found that general types of relationships may effectively mediate analogy item solving.

The verbal analogy item has long been considered the best single type of item to measure general intelligence. The practical importance of the verbal analogy item is attested to by the popularity with which it is used in selecting individuals for special educational curriculums, higher education, and training from a variety of ability tests (e.g., Scholastic Aptitude Test, School and College Ability Test, Miller Analogy Test). However, during the last several years, the lack of firm knowledge about the nature of what is measured by general intelligence tests (Ebel, 1973; McNemar, 1964) has caused increasing discontent among psychologists and educators. The persistence of such doubts, despite the wealth of correlational validity data, has indicated to some methodologists (McNemar, 1964; Messick, 1972) the need for experimental data on the psychological processes used to solve intelligence test items.

The nature of what is measured by intelligence tests is an especially pressing contemporary problem because of claims that the tests discriminate unfairly against certain subcultural groups. The basic issue is whether intelligence test performance may be meaningfully compared between persons of different subcultural groups when their learning experiences may be quite different. As Holtzman (1971) has noted, one of the most volatile criticisms leveled against intelligence testing is that the tests measure the values of the dominant culture rather than intellectual capacity or potential. Although the use of capacity tests which are blind to both situational and cultural factors has been increasingly criticized by major theorists (e.g., Cole & Bruner, 1971), classical models of ability tests do not permit the incorporation of varied learning experiences into the estimation of intellectual potential. Thus, a new basic model of ability and a new type of validity data are needed to answer pressing concerns about the nature of what is measured by intelligence tests.

Cognitive learning treatments on test-related material may potentially be used to investigate both the cognitive processes underlying test performance and the possible incorporation of learning experiences into ability estimates. Popular experimental methods for studying cognitive processes in verbal learning have either supplied mediators (Rohwer, 1971) or instructed subjects to use strategies (Paivio & Foth, 1970) that

[1] This study was supported by an Office of Naval Research Contract No. N00014-68-A-00141-0003 to René V. Dawis through the Center for the Study of Organizational Performance and Human Effectiveness at the University of Minnesota.

[2] Requests for reprints should be sent to Susan E. Whitely, who is now at the Psychology Department, University of Kansas, Lawrence, Kansas 66045.

may potentially facilitate performance. Intervention may be used in a similar way to study psychological processes in solving analogy items, by comparing the facilitation of item performance when supplying mediators or when instructing subjects to use specified strategies to solve items.

Similarly, cognitive learning treatments may also potentially be used to correct ability estimates for varied learning experiences. Whitely and Dawis (in press) explicated an aptitude–ability model of testing from a test–intervention–retest sequence. A basic assumption for the Whitely and Dawis model is that test score gain over intervention mirrors the effects of deficient previous learning experiences. The aptitude–ability model uses a weighted combination of two scores, initial ability and test gain, to estimate true potential (aptitude). Test gain functions as a suppressor variable in the model, which increases the correlation with true potential by partialing out some of the experience-specific variance from the initial ability measure.

The current study has the dual purpose of investigating (*a*) the psychological processes in analogy test performance and (*b*) the modifiability of latent ability scores. Both purposes may be accomplished by experimentally studying the effects of different learning treatments on test performance. The magnitude and nature of the effects of the learning treatments are specially examined for potential standardization into a test–intervention–retest sequence.

Method

Design and Subjects

The subjects were 184 students randomly selected from class lists within two inner-city high schools located in St. Paul, Minnesota. The selected students were randomly assigned to one of six conditions. Each condition was administered once in each school, with the group size varying from 8 to 20 students each. The one-and-a-half-hour session consisted of an analogy item pretest, a treatment condition (or "filler" task), and an analogy item posttest.

Test Materials

One persistent problem in the measurement of change is the equivalency of test forms. For tests developed according to traditional measurement models, test means, variances, and intercorrelations must be equal for difference scores to be meaningful. However, with the newer logistic models of latent traits (i.e., Rasch, 1966), statistically equivalent forms may be obtained from any two subsets if item parameters are scaled and anchored according to the procedure described by Wright and Panchapakesan (1969). Furthermore, difference scores between forms may be individually corrected for measurement error by computing a standardized difference score. Individuals with low-probability standardized differences (a *z* ratio computed by dividing the differences in Rasch ability scores between forms by the measurement error associated with each score) vary in ability more than can be expected from measurement error alone. The standardized difference score permits greater objectivity in the measurement of ability change as compared to traditional ability difference scores, since test-form differences in difficulty and precision at various score levels are statistically controlled.

A set of 142 verbal analogy items was developed to fit the Rasch logistic model of test performance. A prime feature of the analogy items was an unusually low vocabulary level (Lorge-Thorndike frequency appropriate for sixth grade level or less), so that item solving would maximally depend on subjects' ability to educe relationships rather than on knowledge of words. A combined sample of 2,517 students were used to calibrate and anchor the items according to the procedure outlined by Wright and Panchapakesan (1969). The 142 items were divided into the following mutually exclusive sets for the experimental materials: (*a*) a 41-item analogy pretest, (*b*) a 41-item analogy posttest, (*c*) 50 items for the intervention, and (*d*) 10 instructional examples. Although not required for equivalent test forms with items calibrated by the Wright and Panchapakesan (1969) procedure, items were matched for difficulty between forms to allow raw score comparisons and to equate measurement precision at the various ability levels.

Experimental Conditions

The experimental conditions consisted of five interventions and a control. In the control condition, three tests[3] which are not highly correlated with analogy item performance were given as a "filler" task. The experimental conditions focused on practicing item solving, and the 50 intervention items were administered in a fixed order for all five experimental conditions. All items were exposed on a screen for 25 seconds each (10 seconds extra were given for feedback conditions) by a carousel slide projector. For two of the experimental conditions, the 50 items were administered as a practice task only. In one practice condition, item

[3] These tests were the Minnesota Clerical Test and two tests from the French Kit of Cognitive Reference Tests, the Apparatus Test and the Surface Development Test.

feedback was provided by a special answer booklet which displayed the correct answer after the subject had completed an item (P+). In the other practice condition, no feedback was given (P).

In the other three experimental conditions, several units of instruction were interjected between subsets of the 50 intervention items. Although the content of the instruction was constant, both the use of structural aids (described below) and feedback varied between the three conditions with instruction. The three conditions were defined as follows: feedback but no structures (I+); structures but no feedback (I_s); feedback and structures (I_s+). In the conditions with feedback, I_s+ and I+, the correct answer was displayed in a special answer booklet, as in condition P+.

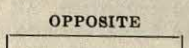
Instructional Units

The instruction was based mainly on Whitely's (1973) empirical study of the types of analogy relationships designated by a group of successful test performers to describe the word-pair relationship.[4] In the Whitely (1973) study, subjects sorted 60 items into relationally homogeneous types according to their own implicit categories. Whitely (1973) identified the following eight latent categories of relationships from a latent partition analysis (Wiley, 1967) of the college students' categorizations: (*a*) similarities, (*b*) opposites, (*c*) word pattern, (*d*) class membership, (*e*) class naming, (*f*) conversion, (*g*) functional, and (*h*) quantity.

All but two of the eight instructional units in the current study were designed to teach subjects to use the eight analogy relationships in solving items by presenting the type names and corresponding definitions followed by examples for each type. Each of the first six units were followed by a subset of the intervention items, with the types of relationships in the items varying according to the instructional unit (see Table 5 for set content).

The six units on the relationship types were followed by two additional units of instruction for which no previous research was available. The seventh unit concerned the position of the related pair and blank in the analogy item stem, while the eighth unit of instruction concerned the role of response alternatives.

Two types of structure were supplied for intervention items in I_s+ and I_s. An example of a fully structured analogy item is presented in which the type of relationship is given and the arrows point to the related pair:



Eat:Slow ::___: Fast

1) Speed 2) Drive 3) Fast 4) Drink 5) Ate

(Answer: 3)

Both types of structure were given in the first four item subsets (arrows and relationship), while the remaining items had, at most, only one type of structure (see Table 5).

Results

Two types of results were obtained. Data from pretest–posttest scores are relevant to the modifiability of ability estimates over various types of intervention. Data on the intervention items are more relevant to psychological process, since the concurrent effects of intervention can be analyzed.

Analogy Test Scores

The measurement of change has been so controversial that some methodologists have recommended not using gain scores (Cronbach & Furby, 1970). However, gain has an irreplaceable role in the Whitely and Dawis (in press) model as a suppressor variable. Because of the controversy, several different variables which measure treatment effects were analyzed, including raw gain, Rasch gain, and the residualized posttest measure recommended by Cronbach and Furby (1970). Standard differences, a special feature of Rasch-scaled test forms, were also analyzed to compare change scores to measurement error. The means and standard deviations for all six dependent variables are presented in Table 1.

Dunnett's *t* statistic was used to detect the significance of the interventions to change latent ability scores, since the level of significance is protected over the set of comparisons of each experimental group to the no-intervention control. Table 2 presents the *t* values for Dunnett's statistic, comparing the analogy test means of each experimental group to the control group. None of the experimental groups differed from the control group on either the raw score or Rasch ability score derived from the analogy pretest. On the change measures, raw gain score, Rasch gain score, and residualized posttest score, the only experimental group to perform significantly better than the control was I_s+ (instruction with feedback and structure). Practice with feedback, practice without feedback, instruction with feedback only, and instruction with

[4] The complete instructions are available upon request from René V. Dawis, Psychology Department, University of Minnesota, Minneapolis, Minnesota 55455.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF ANALOGY TEST SCORES FOR THE FIVE EXPERIMENTAL CONDITIONS AND THE CONTROL GROUP

Condition	*n*	Analogy Test score[a]									
		Raw pretest		Raw gain		Rasch pretest		Rasch gain		Residualized posttest	
		$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
C	31	25.10	9.04	2.29	5.15	−.33	1.25	.35	.82	−1.46	5.05
P	32	25.19	9.24	3.53	3.92	−.17	1.28	.58	.57	−.21	3.71
P+	36	24.17	7.99	3.52	4.89	−.32	1.03	.74	.86	−.31	5.01
I+	26	20.31	9.83	3.38	5.86	−.77	1.43	.63	.75	−.84	5.76
I_s+	34	23.62	9.31	6.26	5.47	−.36	1.28	1.04	.73	2.37	5.14
I_s	25	24.20	7.91	4.00	4.51	−.15	1.10	.71	.67	.16	4.60

Note. Abbreviations: C = control, P = practice, P+ = practice with feedback, I+ = instruction with feedback, I_s+ = instruction with structure and feedback, and I_s = instruction with structure.

[a] The *F* values for Bartlett-Box test of equality of variance are as follows: raw pretest = .43, raw gain = 1.11, Rasch pretest = .80, Rasch gain = 1.29, and residualized posttest = 1.71.

structure only failed to yield significantly higher change scores than the no-intervention controls.

Table 3 presents the means, standard deviations, and range of the standardized difference scores for individuals in each group. The average of the Rasch gain between analogy tests exceeded the measurement errors associated with each score only in the instruction condition with both feedback and structure ($\bar{z} > 1.65, p < .05$, one-tailed).

Table 4 presents the correlations of pretest scores, gain scores, and posttest scores. The pretest correlates highly with the posttest in all conditions for both raw and Rasch scores, indicating that subjects' relative ability does not change drastically. The raw gain score correlated significantly with pretest scores in only two conditions, P and I_s+. Since McNemar (1969) has demonstrated that measurement error may produce significant negative correlations between gain and pretest scores, Rasch gain scores corrected for measurement error (standardized difference scores) were correlated with the Rasch scores on the pretest. The significant negative correlation in I_s+ and P (shown on Table 4) indicate that the correlations may not be attributed to measurement error.

Intervention Items

To compare the treatment effects on the intervention items, two special features of the data had to be considered: (*a*) The items are highly intercorrelated; and (*b*) the intervention units dealt separately with

TABLE 2

DUNNETT *t* VALUES FOR ANALOGY TEST MEAN SCORES OF FIVE EXPERIMENTAL CONDITIONS AS COMPARED TO A CONTROL GROUP

Experimental condition	Score				
	Raw pretest	Raw gain	Rasch pretest	Rasch gain	Residualized posttest
P	−.01	1.19	.11	1.05	1.23
P+	−.25	1.03	−.41	1.64	.99
I+	−1.74	.77	−1.57	1.05	.49
I_s+	−.48	3.01*	−.50	3.38*	3.06*
I_s	−.22	1.35	−.31	1.51	1.31

Note. Abbreviations: P = practice, P+ practice with feedback, I+ = instruction with feedback, I_s+ = instruction with structure and feedback, and I_s = instruction with structure.

* $p = .01$.

TABLE 3

MEANS, STANDARD DEVIATIONS, AND RANGE OF STANDARDIZED DIFFERENCE SCORES IN SIX CONDITIONS

Group	Standardized differences			
	$\bar{X}$	*SD*	Maximum	Minimum
P	1.10	1.00	2.84	−.83
P+	1.10	1.18	3.91	−.93
I+	1.09	1.40	4.30	−2.58
I_s+	1.75	1.23	4.65	−.23
I_s	1.22	1.13	3.30	−.93
C	.75	1.32	3.22	−2.10

Note. Abbreviations: P = practice, P+ = practice with feedback, I+ = instruction with feedback, I_s+ = instruction with structure and feedback, I_s = instruction with structure, and C = control.

the various types of relationships which vary from item to item. Therefore, it was desirable to retain the items as separate dependent variables in the item subsets but, at the same time, to utilize the item intercorrelations in the treatment comparison statistic. A multivariate analysis of variance (MANOVA) most adequately meets these requirements, although the results are somewhat biased by the use of dichotomous variables. Table 5 presents the *F* values from the MANOVA on each dependent variable vector. The following set of orthogonal contrasts were computed:

Effect	P	P+	I+	I_s+	I_s
1. Instruction vs. practice	−3	−3	+2	+2	+2
2. Feedback in practice	−1	+1			
3. Feedback in instruction			+1	+1	−2
4. Structure in instruction			−1	+1	

With the total set of 50 items as the dependent variable vector, Table 5 shows that instruction yielded significantly better performance than practice. Similarly, feedback facilitated performance in both instruction and practice conditions while item structure did not have an overall effect on item performance. With the individual item subsets as dependent variable vectors, Table 5 shows that instructional groups have superior performance on all subsets except Item Set 2, class membership, class naming, and conversion relationships. Similarly, feedback was associated with significantly higher probabilities of solving items on all subsets for instructional groups, with the exception of Item Set 1. However, feedback facilitated performance irregularly over item subsets in the practice conditions.

A group which received structured items (I_s+) performed better than the nonstructured group (I+) on only three item subsets. All the significant subsets had the

TABLE 4

CORRELATIONS OF ANALOGY TEST SCORES IN SIX EXPERIMENTAL CONDITIONS

Variable	Condition					
	C	P	P+	I+	I_s+	I_s
n	31	32	36	26	34	25
Pretest with posttest, raw score	.84**	.91**	.86**	.83**	.81**	.87**
Pretest with gain, raw score	−.20	−.35*	.07	−.19	−.43**	.03
Pretest with posttest, Rausch scores	.79**	.90**	.83**	.87**	.84**	.88**
Pretest with standardized difference, Rausch scores	−.26	−.38**	.10	−.20	−.42**	.04

Note. Abbreviations: C = control, P = practice, P+ = practice with feedback, I+ = instruction with feedback, I_s+ = instruction with structure and feedback, and I_s = instruction with structure.

* *p* significant at .05.

** *p* significant at .01.

TABLE 5

F VALUES FOR MULTIVARIATE ANALYSIS OF VARIANCE OF INTERVENTION, TOTAL, AND SUBSET VECTORS

Item set	No. items	Structure supplied with items[a]	*df*	Contrast effect			
				Practice vs. instruction	Feedback in practice	Feedback in instruction	Structure
1. Similarities, opposites, and word pattern relationships	5	RP	5/143	3.37**	1.24	2.16	2.29*
2. Class membership, class naming, and change into relationships	5	RP	5/143	.73	2.26*	3.46**	1.36
3. Functional and quality relationships	5	RP	5/143	3.13**	2.09	5.92**	2.46*
4. Mixed relationships	10	RP	10/138	7.91**	1.64	3.56**	2.07*
5. Mixed relationships	5	P	5/143	5.15**	2.43*	5.75**	1.24
6. Unidentified relationships	5	P	5/143	2.53*	1.95	2.42*	.79
7. Mixed relationships	10	R	10/138	1.90*	2.43**	5.75**	.84
8. Mixed relationships	5	—	5/143	4.27**	3.31**	13.24**	.79
Total	50	—	50/98	3.65**	1.61*	2.66**	1.01

[a] Abbreviations: R = type of relationship is supplied and P = related pair is supplied.
* p significant at .05.
** p significant at .01.

full structure supplied, that is, both type of relationship and related pairs. The only full structure subset which did not favor I_s+, Item Set 2, was also the only subset for which instructional groups did not score higher than practice groups. Partial structure, in Item Sets 5, 6, and 8, was associated with no significant differences. Item Set 8 had no structure for any group and did not vary significantly between groups.

DISCUSSION

The results indicated that experimental intervention can *sometimes* significantly increase individual ability scores, as estimated from performance on verbal analogy items. Pretest to posttest comparisons of three change measures—raw gain scores, Rasch-scaled gain score, and residualized posttest score—yielded significant ability increases in only one of the five intervention conditions, instruction with both feedback and structure. Practice on the intervention analogy items, even when accompanied by feedback, had no significant effect in changing ability scores. Similarly, neither instruction with feedback alone, nor instruction with structural aids alone, led to ability changes.

The finding that only the most intensive intervention condition significantly changes ability scores may be more readily understood in comparison to previous attempts to study ability change over intervention. The British studies on the effects of coaching and practice on ability test scores, summarized by Vernon (1954), have been the most extensive research findings on ability as modified by test-related interventions. The British researchers were concerned with norm validity when the test preparation given to students varied between schools, and they devised a series of test-related instructional units which were given to all students over several days. Similiarly, Jacobs and Vandeventer (1969) in the United States were able to modify Raven's Progressive Matrices scores by a lengthy series of interventions. In contrast to these studies, the intervention in the present study was quite brief (a single 50-minute session), and the failure for the less intensive conditions to change ability scores may have resulted from the severe time restrictions.

The results obtained also have some implications from the use of modifiability in ability measurement, such as in the aptitude–ability model (Whitely & Dawis, in press). First, ability scores can be significantly altered by an intervention which is short enough to standardize into a single testing session. Second, significant test gain may potentially function as a suppressor variable, as shown by its negative correlation with initial ability, even when corrected for measurement errors. These findings support the feasibility of the aptitude–ability model in estimating true potential from a weighted combination of initial ability and test gain obtained in a test–intervention–retest sequence.

Some clues as to the psychological processes involved in solving analogy items may be gleaned from the analysis of the intervention items as well as from the pretest–posttest results. Cognitively oriented instruction, designed to teach the taxonomy of types of analogy relationships described by a test-sophisticated group, had an immediate facilitatory effect on analogy item solving. The facilitation of analogy item performance by general types of relationships is contrary to expectation if item solving depends on a simple associational process, as has been indicated by some previous research (Gentile, Kessler, & Gentile, 1969; Willner, 1964). Gentile et al. (1969) had suggested that the relationships used to solve analogy items were specific to both persons and items. The current study suggests that the use of an empirically derived relational taxonomy in item solving facilitates performance. Apparently a more complex process than simple association underlies analogy item solving.

Interestingly, instruction was not effective for all types of relationships. Instruction significantly increased item performance for similarities, opposites, word pattern, functional and quantity relationships, while instruction in class membership, class naming, and conversion relationships had no apparent facilitatory effect on analogy item solving. Unfortunately, from the design in the current study, it is impossible to separate the adequacy of the definition of ineffective relationships from the degree to which special instruction in the relationship actually facilitates performance.

Supplying cognitive mediators (i.e., type of relationship, related pair) with analogy items effectively facilitated item solving only when full structure was provided. Supplying type of relationship or related pair did not alone significantly improve item performance. Apparently the mediator must elicit the full association rather than simply reducing the associative possibilities. With respect to Spearman's (1923) three principles of intelligence and cognition, when the full structure is provided, the task for the subject is to educe a correlate for the remaining stem word by using the supplied relationship. However, when only partial structure is supplied, the subject still has the dual task of educing relationships and correlates. This is similar to the initial task, except that the associative possibilities would theoretically be reduced.

In summary, this study finds that modifiability of latent ability scores depends on type of intervention and that, minimally, a complex associational process is needed to account for the facilitation of analogy items by general types of relationships.

REFERENCES

Cole, M., & Bruner, J. S. Cultural differences and inferences about psychological processes. *American Psychologist*, 1971, **26**, 867–876.

Cronbach, L. J., & Furby, L. How we should measure change—or should we? *Psychological Bulletin*, 1970, **74**, 68–80.

Ebel, R. The future of the testing of abilities. Presidential address to the Annual Meeting of the American Educational Research Association, New Orleans, March 1973.

Gentile, J. R., Kessler, D. K., & Gentile, P. K. Process of solving analogies items. *Journal of Educational Psychology*, 1969, **60**, 494–502.

Holtzman, W. H. The changing world of mental measurement and its social significance. *American Psychologist*, 1971, **26**, 546–553.

Jacobs, P. I., & Vandeventer, M. *The learning and transfer of double-classification skills: A replication and extension.* (ETS Research Bulletin 69-88) Princeton, N.J.: Educational Testing Service, 1969.

McNemar, Q. Lost: Our intelligence? Why? Address of the President to the Seventy-Second Annual Convention of the American Psychological Association, Los Angeles, Sptember 1964.

McNemar, Q. *Psychological statistics*. New York: Wiley, 1969.

Messick, S. Beyond structure: In search of functional models of psychological process. *Psychometrika*, 1972, **37**, 357–375.

Paivio, A., & Foth, D. Imaginal and verbal mediators and noun concreteness in paired-associate learning: The elusive interaction. *Journal of Verbal Learning and Verbal Behavior*, 1970, **9**, 384–390.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, **19**, 49–57.

Rohwer, W. D., Jr. Prime time for education: Early childhood or adolescence? *Harvard Educational Review*, 1971, **41**, 319–341.

Spearman, C. *The nature of intelligence and the principles of cognition*. London: Macmillan, 1923.

Tinsley, H. E. A. An investigation of the Rasch simple logistic model for tests of intelligence or attainment. Unpublished doctoral dissertation, University of Minnesota, 1971.

Vernon, P. E. Symposium on the effects of coaching and practice in intelligence tests. *British Journal of Educational Psychology*, 1954, **24**, 57–63.

Whitely, S. E. Types of relationships in reasoning by analogy. Unpublished doctoral dissertation, University of Minnesota, 1973.

Whitely, S. E., & Dawis, R. V. A model for psychometrically distinguishing aptitude from ability. *Educational and Psychological Measurement, in* press.

Wiley, D. E. Latent partition analysis. *Psychometrika*, 1967, **32**, 183–193.

Willner, A. An experimental analysis of analogical reasoning. *Psychological Reports*, 1964, **15**, 479–494.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

Wiseman, S., & Wrigley, J. The comparative effects of coaching and practice on the results of verbal intelligence tests. *British Journal of Psychology*, 1953, **44**, 83–94.

Wright, B., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, **29**, 23–48.

(Received October 15, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 718-725

AUDITORY-VISUAL INTEGRATION AND READING PERFORMANCE IN LOWER-SOCIAL-CLASS CHILDREN

GERALD W. JORGENSON[1]
Case Western Reserve University

ELIZABETH M. HYDE
John Carroll University

The present study examined the relationship of auditory–visual integration and reading performance in a sample on first- and second-grade subjects from primarily lower-class backgrounds. The effects of grade level, sex differences, IQ, and single-modality auditory and visual functioning were examined as possible moderating variables. The results suggested that auditory–visual integration ability in this sample was developmental; however, the subjects functioned at a lower level than would be expected of middle-class subjects. Significant correlations were obtained between auditory–visual integration and reading vocabulary in the total sample and in the second-grade boys. Although neither IQ nor single-modality ability was a significant contributor to this relationship, the results suggested that grade level, sex differences, and socioeconomic status did make significant contributions and should be considered when interpreting auditory–visual integration research.

Since the original work of Birch and Belmont (1964, 1965), studies have demonstrated that a significant relationship exists between reading performance and the ability to judge the equivalence of auditory and visual stimuli (Beery, 1967; Ford, 1967; Kahn & Birch, 1968; Muehl & Kremenak, 1966; Reilly, 1971; Sterritt & Rudnick, 1966). With the possible exception of Reilly (1971), these studies have included primarily middle-class subjects. The present study examined the relationship of auditory–visual integration and reading performance in primarily lower-class first- and second-grade subjects to determine whether the conclusions of previous studies apply to this population. Within this general framework, four additional variables which may influence the relationship between auditory–visual integration skill and reading performance were investigated: grade-level placement, sex-group membership, IQ, and single-modality functioning. The rationale for studying each of these variables is presented below.

[1] Requests for reprints should be sent to Gerald W. Jorgenson, Department of Education, Case Western Reserve University, Cleveland, Ohio 44106.

Grade-Level Placement

Birch and Belmont (1965) established that auditory–visual integration ability increased most rapidly in children between kindergarten and Grade 2. Kindergarten children functioned at little better than a chance level, while an average second grader got 80% of the items on a 10-item test correct. The authors found a correlation of .70 ($N = 29, p < .001$) between auditory–visual integration and reading performance in Grade 1. At higher grade levels, the correlations were not significant. In addition, at each grade level except kindergarten and sixth grade, a significant positive relationship existed between auditory–visual integration and IQ, with a range of coefficients from .34 to .57.

Subsequent research (Kahn & Birch, 1968), utilizing an expanded and more difficult auditory–visual integration test instrument, has indicated that auditory–visual integration continues to develop throughout the elementary school years and continues to be significantly related to reading performance at least through the sixth grade. However, the conclusion of Birch and Belmont

that auditory–visual integration ability develops most rapidly between kindergarten and Grade 2 remains viable. Precisely when auditory–visual integration "peaks" during this time span, and when the ability is most related to reading performance, remains unclear.

Reilly (1971) found significant differences in auditory–visual integration performance between first- and second-grade subjects, concluding that better discrimination is associated with older children. In first-grade subjects he found a correlation of .23 ($N = 60$, $p < .05$) between auditory–visual integration and vocabulary as measured by the Gates-MacGinitie Reading Test. Auditory–visual integration was not significantly related to either reading comprehension or total reading score. In the second grade, all correlations were significant: auditory–visual integration and vocabulary ($N = 56$, $r = .65$, $p < .005$), comprehension ($N = 56$, $r = .71$, $p < .005$), total reading score ($N = 56$, $r = .70$, $p < .005$). Since auditory–visual integration accounted for 42% of the reading vocabulary variance and 50% of the reading comprehension variance in the second-grade subjects, Reilly concluded that it was at that level that auditory–visual integration was most powerfully related to reading ability. This conclusion is the opposite of the one suggested by Birch and Belmont, who found that auditory–visual integration accounted for 49% of the reading variance in first-grade subjects and 18% of the variance in second graders. Although these results could perhaps be explained by methodological differences in the two studies—Reilly used a 20-item auditory–visual integration test, rather than the 10-item Birch and Belmont test, as well as a different reading performance measure—the differing conclusions led the present authors to examine the effect of grade-level placement on auditory–visual integration performance level and on the magnitude of the relationships between auditory–visual integration and reading performance.

Sex Differences

Muehl and Kremenak (1966) found that the sex of the subjects included in their sample of first-grade children was not significantly related to auditory–visual integration performance. Reilly (1971), on the other hand, found that the relationship of auditory–visual integration to reading performance at different grade levels was affected by the sex of the subjects. For example, in first-grade boys, no significant correlation was found between auditory–visual integration and reading performance; in first-grade girls, a significant correlation existed between auditory–visual integration and reading vocabulary. In the present study, sex differences were examined as possible moderating variables to see if Reilly's results could be replicated.

IQ

The relationship between auditory–visual integration and intelligence has been examined in previous studies but with contradictory results. Birch and Belmont (1964, 1965), Sterritt and Rudnick (1966), Muehl and Kremenak (1966), and Beery (1967) are among those who have found that auditory–visual integration significantly predicted reading performance, independently of the influence of IQ. Kahn and Birch (1968) found that controlling the effects of IQ did not effect the significant correlation of auditory–visual integration and word knowledge but reduced the correlation of auditory–visual integration and comprehension in Grade 1 to an insignificant level. Ford (1967), however, found that controlling for the effects of IQ reduced the level of correlation between auditory–visual integration and all measures of reading performance to an insignificant level. Thus the question of whether auditory–visual integration is a significant predictor of reading performance when the effects of general intelligence are controlled remains unresolved.

Single-Modality Ability

There has been relatively little examination in the literature of the effects of auditory and/or visual functioning on integrative ability. In terms of the effect of auditory functioning on auditory–visual integration performance, three studies (Birch & Belmont, 1964; Ford, 1967; Kahn & Birch,

1968) administered the digit span subtest of the Wechsler Intelligence Scale for Children (WISC) to all of a sample of the subjects. None of these studies found a significant relationship between short-term auditory memory and auditory–visual integration performance. Kahn & Birch (1968) added an additional auditory test, asking the 10 highest and 10 lowest auditory–visual integration scorers at each grade level tested to "sing" or "tap out" patterns which corresponded to the auditory component of the auditory–visual test. This measure did not discriminate between high- and low-auditory–visual scorers, and the authors concluded that single-modality auditory performance did not contribute significantly to integrative performance.

The influence of visual functioning on auditory–visual integration performance has been investigated by Kahn and Birch (1968). As a measure of visual perception, the 10 highest and 10 lowest auditory–visual integration scorers at each grade level tested in the study were asked to choose, from among three choices, the visual dot pattern that matched a visual dot pattern cue. There were no significant differences between low- and high-auditory–visual scorers on this test, and the authors concluded that single-modality visual perception did not make a significant contribution to integration performance.

These conclusions regarding the influence of single-modality abilities must be tentative. Several of the studies employed extreme group procedures wherein samples were split into the highest and lowest auditory–visual integration scorers. Such procedures tend to attenuate relationships which might have emerged had the total sample of subjects been considered. Moreover, the nature of the single-modality tasks utilized in previous studies is questionable. Particularly in the Kahn and Birch study, the visual perception test was too easy, in that only one subject out of the total sample tested made more than two errors. In this same study, one of the auditory as well as the visual single-modality measures consisted of original tests, created exclusively for the purpose of the study. The use of nonstandardized, original single-modality tests for which there is neither validity nor reliability data is questionable. Finally, even when investigators have employed a standardized measure of single-modality auditory functioning, this measure has typically been the WISC digit span—a subtest of general intelligence test whose validity is derived from its contribution to a composite IQ score, not from empirical evidence that it measures auditory memory.

The present study investigated the relationship of auditory–visual integration and single-modality auditory and visual functioning using the Visual Sequential Memory and the Auditory Sequential Memory subtests of the Illinois Test of Psycholinguistic Abilities (ITPA). The ITPA was selected because of its previous use with subjects demographically similar to those in the present study (Stephenson & Gay, 1972; Weaver & Weaver, 1967) and because of its widespread use as a differential diagnostic instrument specifically designed to measure single-modality strengths and weaknesses.

In summary, the present study examined the applicability of previous research findings concerning the relationship between auditory–visual integration and reading performance in a sample of first- and second-grade lower-socioeconomic-status children. In this investigation, the effects of grade level, sex, IQ, and single-modality functioning were investigated as potential variables moderating the relationship between auditory–visual integration and reading skill.

Method

Subjects

Approximately one half of the first- and second-grade student population was randomly selected from an elementary school in a low-socioeconomic-status community near a large midwestern city. This procedure resulted in a sample of 86 students, 52% of whom were in Grade 1 and 48% in Grade 2. Of the first-grade subjects, 19 were females and 26 were males; at the second-grade level, 22 were females and 19 were males.

Of the total sample, 32% were A.D.C. welfare recipients, and an additional 34% were from families who fell below the poverty level established by the federal government.

Test Instruments

Auditory–visual integration performance was assessed using the 10-item instrument and testing procedures developed by Birch and Belmont (1965). The examiner presented an auditory cue to the subject by tapping on a desk with a pencil in a predetermined sequence and then exposed three separate dot patterns. The subjects selected which of the three dot patterns matched the auditory cue. To avoid the possibility of the subjects receiving visual cues (Sterritt & Rudnick, 1966), the examiner sat to the side of and slightly behind each subject to present auditory cues. Except for this change, the methods of test administration and the directions that were given followed those outlined by Birch and Belmont (1964, 1965). The method included a demonstration of the auditory and visual components of the auditory–visual integration task, the presentation of three practice items, followed by the presentation of the 10 test items.

Auditory memory and visual memory were measured by the Auditory Sequential Memory subtests of the revised Illinois Test of Psycholinguistic Abilities (Kirk, McCarthy, & Kirk, 1968).

All subjects received the appropriate form of the Gates-MacGinitie Reading Test: Primary A for Grade 1 and Primary B for Grade 2. These tests yielded vocabulary and comprehension scores which were transformed into standard scores for purposes of analysis.

In addition to the above information, Kuhlmann-Anderson IQ scores were obtained for each subject. In the school district from which the sample was selected, this test is routinely administered in the spring semester during the first-grade year. The results were available from the school records.

Procedure

The auditory–visual integration test was administered to all subjects in individual testing sessions by the first author using the methods described above. Auditory memory and visual tests were administered to all subjects in individual testing sessions by the second author, using the standardized directions for administration specified in the ITPA manual. All testing was done during the morning hours of the school day.

The Gates-MacGinitie Reading Tests were administered by the subjects' classroom teachers who used the procedures outlined in the test manuals.

Results and Discussion

Since the auditory–visual integration instrument and procedures were patterned after those developed by Birch and Belmont (1965), Table 1 presents means, standard deviations, and ranges of auditory–visual integration scores for the Birch and Belmont primarily middle-class sample, in comparison with similar information for the subjects in the present study. The auditory–visual integration task appears to be more difficult for the subjects in the present study. The 4.4 mean for the second-grade subjects is comparable to the 4.1 mean for the kindergarten subjects in the Birch and Belmont study. In addition, 70% of Birch and Belmont's kindergarten subjects got 4 or fewer items correct on the auditory–visual integration test, while at first grade, only 30% were functioning at this level. In the present study, 76% of the first-grade subjects got 4 or fewer correct; the comparable percentage at the second-grade level was 51%. Finally, Birch and Belmont found that second-grade subjects averaged 80% correct on the 10-item test. The average percentage correct in the present study was approximately 45%.

TABLE 1

Comparison of Auditory–Visual Integration Performance in the Present Study with Performances Reported by Birch and Belmont (1965)

Study	Grade	*n*	$\bar{X}$	*SD*	Range
Birch and Belmont	1	30	5.6	2.2	1–9
Present	1	45	3.6	1.3	1–6
Birch and Belmont	2	30	7.9	1.6	5–10
Present	2	41	4.4	1.9	1–8

On the basis of their results, Birch and Belmont concluded that the period of most rapid development of integration skill was between kindergarten and Grade 2. The results of the present study, however, suggest that integration skill has not developed to a level comparable to that of middle-class children by the second grade. This result suggests that the conclusions of Birch and Belmont should be modified to take account of the socioeconomic-status level of the subjects.

Further analysis of the data in the present study involved an examination of the relationship of auditory–visual integration to sex and grade level. Table 2 present means and standard deviations of the total number correct on the auditory–visual integration task for each sex at each grade level tested. These data were subjected to a 2 (Sex) × 2

TABLE 2
MEANS AND STANDARD DEVIATIONS OF AUDITORY–VISUAL INTEGRATION SCORES BY GRADE AND SEX

Grade	*N*		$\bar{X}$		*SD*	
	Boys	Girls	Boys	Girls	Boys	Girls
1	26	19	3.5	3.9	1.0	1.4
2	19	22	4.2	4.6	2.0	1.9

(Grade Level) unweighted means two-way analysis of variance. While neither the interaction effect nor the sex main effect were statistically significant, there was a statistically significant grade level main effect ($F = 4.37$, $df = 1/82$, $p < .05$). Thus, although there are no significant differences between males and females in mean auditory–visual integration scores, auditory–visual integration scores of second graders were significantly higher than those achieved by first graders.

As a further examination of the age relatedness of the auditory–visual integration task, Pearson product moment correlations were computed between auditory–visual integration scores and chronological age. This analysis resulted in a statistically significant correlation coefficient of .304 ($df = 84$, $p < .01$). This result, along with the statistically significant grade-level main effect reported above, suggests that in a predominantly lower-socioeconomic-status sample, auditory–visual integration is a developmental, linearly age-related skill. The low scores on the auditory–visual integration task, however, suggest that the most rapid period of development occurs at some point beyond the second grade.

The effects of auditory and visual short-term memory on auditory–visual integration were investigated by computing Pearson product-moment correlations between auditory–visual integration scores and standard scores obtained on the Auditory Sequential Memory and Visual Sequential Memory subtests of the ITPA. In light of the interest in sex as a moderating variable, correlations were computed separately for each sex as well as for the total sample. Table 3 presents these results.

None of the total sample (combined sexes) correlations were statistically significant. In the separate analysis by sex, the visual memory subtest was significant at the .05 level for the second-grade boys only.

The results for auditory short-term memory substantiate previous research (Birch & Belmont, 1965; Ford, 1967; Kahn & Birch, 1968). It appears that short-term auditory memory, as measured by either the WISC digit span subtest or the Auditory Sequential Memory subtest of the ITPA, is unrelated to auditory–visual integration skill.

Like auditory sequential memory, visual sequential memory was generally not significantly related to auditory–visual integration performance. While the one significant relationship between visual sequential memory and auditory–visual integration in the second-grade boys might have arisen from chance factors, the fact that visual sequential memory accounted for almost

TABLE 3
PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENTS BETWEEN AUDITORY–VISUAL INTEGRATION PERFORMANCE AND AUDITORY AND VISUAL MEMORY

Grade	Boys		Girls		Boys & Girls	
	Auditory	Visual	Auditory	Visual	Auditory	Visual
1	.094[a]	−.113	−.082[c]	.000	−.061	.003
2	.398[b]	.549*	−.259[d]	.111	−.048	.083
1 & 2	.223	.264	−.119	.098	.022	.188

[a] $n = 26$.
[b] $n = 19$.
[c] $n = 19$.
[d] $n = 22$.
* $p < .05$.

one third of the variance of auditory–visual integration in second-grade boys suggests that single-modality visual functioning may be a significant correlate of auditory–visual integration. Additional research is needed to clarify this point.

To investigate the relationship between auditory–visual integration and reading performance, Pearson product-moment correlations were computed between auditory–visual integration scores and standard scores obtained on the Vocabulary and Comprehension substests of the Gates-MacGinitie Reading Tests. Correlations were computed separately for sex as well as for total sample scores. These results are presented in Table 4.

There was no significant relationship between auditory–visual integration and comprehension. On the other hand, the relationship between auditory–visual integration and the vocabulary measure was significant ($p < .05$) for the combined sexes sample. However, further examination reveals a significant correlation for males but a nonsignificant correlation for females. One might hypothesize that the significant correlation found in the total sample may be largely attributable to the strength of the auditory–visual integration vocabulary correlation found in the boys.

A final issue examined in the present study pertained to the influence of general intelligence on the relationship between auditory–visual integration and reading performance. Kuhlmann-Anderson IQ scores were available for 81 subjects or 94% of the present sample. These IQs ranged from 80 to 143, with a mean IQ score of 107.3 and a standard deviation of 12.02.

TABLE 4

PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT BETWEEN AUDITORY-VISUAL INTEGRATION PERFORMANCE AND READING VOCABULARY AND COMPREHENSION

Grade	Boys		Girls		Boys & Girls	
	Vocabulary	Comprehension	Vocabulary	Comprehension	Vocabulary	Comprehension
1	.374[a]	.086	.110[c]	.078	.262	.099
2	.460[b]*	.349	.281[d]	.061	.016	−.111
1 & 2	.419**	.192	.142	.043	.225*	.056

[a] $n = 26$.
[b] $n = 19$.
[c] $n = 19$.
[d] $n = 22$.
* $p < .05$.
** $p < .01$.

Pearson product-moment correlations were computed between auditory–visual integration and Kuhlmann-Anderson IQ scores. Correlations were computed separately for each sex at each grade level as well as across sexes and grade levels. None of these coefficients was significant. Birch and Belmont (1965) reported a significant, positive relationship between auditory–visual integration and IQ in Grades 1 and 2 ($p < .01$ and $p < .05$, respectively); the present study failed to replicate their finding.

The discrepancy between the two studies may be attributable to methodological artifacts—for example, the two studies employed different measures of IQ, and the auditory–visual integration task employed in the Birch & Belmont study failed to provide an adequate basal for the subjects in the present study and thereby restricted the range of the auditory–visual integration measure. On the other hand, it is also possible that while auditory–visual integration and IQ are significantly correlated in middle-socioeconomic-status subjects, the two measures are not significantly related in a predominantly lower-socioeconomic-status population.

In the present study, the correlations between auditory–visual integration and reading vocabulary were significant for the total sample and for the total sample of boys. First-order partial correlations were computed to determine the stability of these correlations when the effects of IQ were controlled. After partialing, the relationship between auditory–visual integration and vocabulary remained significant ($r = .49$, $p < .01$ for the boys; $r = .30$, $p < .05$ for the total sample). In fact, the negative direction of the IQ/auditory–visual integration correlation ($r = -.153$ for total boys; $r = -.079$ for the total sample) caused the final correlation to be higher than it was origi-

nally. These results corroborate conclusions of Birch and Belmont (1964, 1965), Sterritt and Rudnick (1966), Muehl and Kremenak (1966), Beery (1967), and Kahn and Birch (1968); namely, that independently of the influence of IQ, auditory–visual integration is a significant correlate of reading performance.

General Discussion

The results of the present study suggested that the findings of previous research should be viewed as tentative and restricted to the specific type of subjects under study, that is, typically middle-class boys. While the present study itself must be replicated with additional samples from a lower-socioeconomic-status population and must be expanded to include a larger range of grade levels and samples from several socioeconomic-status backgrounds, there is at least preliminary evidence that both sex and socioeconomic background might be important factors to consider in interpreting not only auditory–visual integration functioning but also the relationship between auditory–visual integration skill and reading performance.

The significant relationship of auditory–visual integration and reading vocabulary in the present sample of boys but not in the sample of girls substantiated Reilly's (1971) suggestion that sex differences should be taken into account in interpreting test results. The potential importance of sex differences is particularly pronounced in the area of auditory–visual integration research since many of the studies conducted (see Birch & Belmont, 1964; Ford, 1967; Kahn & Birch, 1968; Sterritt & Rudnick, 1966; Rudnick, Sterritt, & Flax, 1967) have included only boys in their samples. The present findings, coupled with those of Reilly (1971), suggest that the results of these earlier studies may be limited only to the subjects employed, namely, middle-class boys.

Unlike previous studies, the present study investigated the relationship between auditory–visual integration and single-modality functioning by employing standardized instruments specifically designed to assess auditory and visual memory abilities. The present study supported earlier research (Birch & Belmont, 1946; Ford, 1967; Kahn & Birch, 1968) and found no significant relationship between auditory–visual integration and auditory sequential memory. By contrast, the visual sequential memory task was significantly related to auditory–visual integration performance in the present sample of second-grade boys ($p < .05$). Since the significant relationship obtained in the present study was for second-grade boys only, the result must be viewed as tentative and must be investigated further.

As the literature was being examined in preparation for the present research and as the study was being carried out, it was apparent that an obvious need exists for greater standardization and specificity in auditory–visual integration research. Different measures of reading performance, IQ, and single-modality functioning have been used almost indiscriminately. Of greater import, however, may be the types of auditory–visual integration measures employed. These range from the 10-item matching format using dot–dash stimuli employed by Muehl and Kremenak (1966), to the 10-item multiple-choice format developed by Birch and Belmont (1964, 1965) and utilized in the present study, to the 20-item multiple-choice formats used by Kahn and Birch (1968) and Reilly (1971). A major difficulty, and possibly a cause of much of the conflicting evidence, may be the lack of a standardized auditory–visual integration test instrument that meets basic requirements of reliability and validity. In the estimation of the present researchers, the development of such an instrument would be a fruitful next step.

REFERENCES

Beery, J. W. Matching of auditory and visual stimuli in average and retarded readers. *Child Development*, 1967, **38**, 827–831.

Birch, H., & Belmont, L. Auditory-visual integration in normal and retarded readers. *American Journal of Orthopsychiatry*, 1964, **34**, 852–861.

Birch, H., & Belmont, L. Auditory-visual integragration, intelligence, and reading ability in school children. *Perceptual and Motor Skills*, 1965, **20**, 295–305.

Ford, M. P. Auditory-visual and tactual-visual integration in relation to reading ability. *Perceptual and Motor Skills*, 1967, **24**, 831–841.

Kahn, D., & Birch, H. Development of auditory-visual integration and reading achievement. *Perceptual and Motor Skills,* 1968, **27,** 459–468.

Kirk, S. A., McCarthy, J. J., & Kirk, W. D. *Illinois Test of Psycholinguistic Abilities.* (Revised Ed.) Urbana: University of Illinois Press, 1968.

Muehl, S., & Kremenak, S. Ability to match information within and between auditory and visual sense modalities and subsequent reading achievement. *Journal of Educational Psychology,* 1966, **57,** 230–238.

Reilly, D. H. Auditory-visual integration, sex, and reading achievement. *Journal of Educational Psychology,* 1971, **62,** 482–486.

Rudnick, M., Sterritt, G. M., & Flax, M. Auditory and visual rhythm perception and reading ability. *Child Development,* 1967, **38,** 581–588.

Stephenson, B., & Gay, W. Psycholinguistic abilities of black and white children from four SES levels. *Exceptional Children,* 1972, **38,** 705–709.

Sterritt, G., & Rudnick, M. Auditory and visual rhythm perception in relation to reading ability in fourth grade boys. *Perceptual and Motor Skills,* 1966, **22,** 859–864.

Weaver, S., & Weaver, A. Pscholinguistic abilities of culturally deprived Negro children. *American Journal of Mental Deficiency,* 1967, **72,** 190–198.

(Received December 17, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 726–730

IMPACT OF COEDUCATION ON "FEAR OF SUCCESS" IMAGERY EXPRESSED BY MALE AND FEMALE HIGH SCHOOL STUDENTS[1]

RONALD WINCHEL, DIANE FENNER, AND PHILLIP SHAVER[2]

Columbia University

Horner has stated that the motive to avoid success ("fear of success") is increased in females who are forced to compete with males, and she has suggested that coeducation is detrimental to females' academic performance. In the present study, 240 male and female high school seniors were tested for fear of success, some while attending coed and some while attending noncoed high schools. The kind of elementary school the subjects had attended, coed or noncoed, was also recorded. Results indicated that (*a*) both male and female subjects expressed more negative themes when writing stories about a successful female than when writing about a similarly successful male; (*b*) the coed versus noncoed distinction was a potent predictor of fear of success response in female subjects, especially at the elementary school level; and (*c*) when negative consequences of success were mentioned, they were usually social or affiliative if the successful figure in the story was female and nonaffiliative if the figure was male. Theoretical and practical implications were briefly discussed.

Recent studies reported by Horner (1970, 1972) have greatly increased our understanding of the motivational differences between American males and females in achievement situations. It had long been known that a large body of results and inferences concerning achievement motivation in males did not apply for some reason to females (e.g., Atkinson, 1958; French & Lesser, 1964; Lesser, Krawitz, & Packard, 1963; Veroff, Wilcox & Atkinson, 1953), but no adequate explanation had been formulated. Horner hypothesized that females are hampered by a form of anxiety, which she called "fear of success," that is uncommon in males. In order to measure fear of success, Horner asked female college students to tell a story based on the following projective cue: "After first-term finals, Ann finds herself at the top of her medical school class." A majority of her subjects, over 65%, portrayed Ann as anxious or guilty or predicted that Ann's success would have unpleasant consequences, such as loss of femininity or social rejection. In a comparable group of male subjects, responding to a cue describing "John" rather than "Ann" as the successful medical student, less than 10% expressed negative themes. It was also found that females who expressed fear of success on the projective measure performed worse in a competitive than in a noncompetitive situation, whereas most of the males and the minority of females who did not show fear of success performed better under competitive conditions.

Because Horner gave the John cue only to males and the Ann cue only to females, it was impossible to determine whether negative responses were due to sex of subject (suggesting fear and conflict peculiar to females) or to sex of the actor in the cue (suggesting that males and females hold common sex role stereotypes). Horner's studies also left open many developmental

[1] The authors gratefully acknowledge the cooperation of the Yeshiva University High Schools, Brooklyn, and the Joel Braverman Yeshiva High School, Brooklyn.

[2] Requests for reprints should be sent to Phillip Shaver, Department of Psychology, Columbia University, New York, New York 10027.

and educational policy questions, as she herself pointed out (Horner, 1968, Chapter 4). A study by Monahan, Kuhn, and Shaver (1974) has addressed some of these questions. One hundred and twenty 10- to 16-year-old boys and girls responded to the John or Ann cue in a completely crossed design. Both sexes gave more negative responses to the Ann cue, indicating that sex role stereotypes are largely responsible for the "fear of success" imagery. There was also some evidence, however, that motivational differences were being tapped: Females often expressed anxiety and conflict in describing Ann's fate while the males often expressed hostility. The overall frequency of negative responses declined with age, contrary to a trend described by Horner (1972), and more "fear of success" was expressed in response to the John cue than Horner's study would lead one to expect. (This finding has been obtained in other recent investigations as well; see the review by Tresemer, 1973.)

Horner's experimental results (1968, 1970) and subjects' answers to questions concerning competition convinced her that the "motive to avoid success" is aroused most strongly in competitive situations. Considering the prevailing sex roles and the emphasis on education as a means to "get ahead" in a competitive society, she speculated that the developmental trend toward greater fear of success in older college females was due to the increasingly significant competition for valued positions in academic and professional life. Consistent with this is her finding (Horner, 1972) that academically superior female students—those most capable of competing with males for professional positions—are more likely than their less able classmates to show fear of success. In several places, Horner has suggested that such findings may argue against the prevailing emphasis on coeducation, since coeducation increases the salience of cross-sex competition for academic and professional success. This possibility is especially important considering a recently released report from the Carnegie Commission on Higher Education (1973), *Opportunities for Women in Higher Education.* The report indicates that a high proportion of successful women come from noncoed colleges. Although this could be due partly to selective enrollment,

> there are reasons for believing that the experience of attending a women's college is partially responsible. In women's colleges, female students are not reluctant to participate actively in class discussion for fear of losing their feminine appeal in the eyes of male students. They have far greater opportunity to gain experience in leadership roles in campus organizations and activities than women in coeducational institutions, where the top leadership positions nearly always go to men [p. 73].

The report also notes that students in women's colleges are more likely to enter traditionally male fields, such as science, and are more likely to participate in athletics.

The present study was designed to explore the development of fear of success a bit further, with particular emphasis on how such development is affected by coeducation. Male and female high school students from the same geographical area and from similar socioeconomic and cultural backgrounds were given either the Ann or the John cue as a measure of fear of success. Some of them were attending a coed high school and some were not. Moreover, part of each group had attended noncoed elementary schools.[3]

Method

Subjects

Two hundred and fifty-two high school seniors attending Jewish private schools in a middle- to upper-middle-class Brooklyn community completed the fear of success measure in the spring of 1973. After the measure had been taken, each subject was asked what he or she thought the study was about (Horner's work had been discussed in national periodicals that were potentially available to the subjects). Twelve people were eliminated before analyses were performed because they were aware of the purpose of the measure. The number of males and females in each school and cue category is shown in Table 1.

Procedure

Subjects completed the fear of success measure in their classrooms after being assured that

[3] In this paper, "elementary school" refers to Grades 1 through 8.

the results would be seen only by researchers at Columbia University and not by the subjects' teachers. Some received booklets containing the John cue, others received the Ann cue. An attempt was made to equate the number of males and females in each condition, but this was not always successful. All instructions were given in written form. These asked the subject to "write a story" based on the following sentence: After first term finals, Ann (John) finds herself (himself) at the top of her (his) medical school class. "You should tell (*a*) what has happended in the past that led up to this; (*b*) what Ann (John) thinks about this; (*c*) what will happen to Ann (John) in the future." Each subject was given as much time as necessary. After completing a story, each subject answered the question mentioned previously about the probable purpose of the study. All protocols were adequate for scoring and, with the exception of those from the 12 nonnaive subjects, all were included in the analyses ($N = 240$).

Scoring

The initial scoring procedure was identical to Horner's (1970, p. 59). If a protocol expressed an overall positive attitude toward the achievement and toward the actor, it was scored as positive. If it indicated any negative attitudes toward the achievement or the actor or specified negative consequences of the achievement, it was scored as negative. Two raters, one male and one female, scored the protocols independently, agreeing in 96% of the cases. They also divided the negative responses into two categories (agreeing in 93% of the cases): affiliative concerns (e.g., marriage, loss of friends, loss of feminine attractiveness) and nonaffiliative concerns (e.g., monetary considerations, dissatisfaction with work load, claims that the success was due to cheating). All disagreements were easily resolved by discussion.

Results

The percentages of subjects whose protocols were assigned to the "fear of success" category are given in Table 1 for each condition. The results replicate Horner's basic finding: For each sex of subject and type of school category, the proportion of subjects showing fear of success is higher for the Ann cue than it is for the John cue (overall $\chi^2 = 8.71$, $df = 1$, $p < .005$). Of special interest is the significant difference between the proportion of fear of success stories evoked by the Ann cue in females from coed schools and the corresponding proportion for females from noncoed schools (40.9% for the coed school, 15.8% for the noncoed school: $\chi^2 = 4.43$, $df = 1$, $p < .05$; the corresponding percentages for males, 30.8% and 27.3%, are not significantly different).

TABLE 1

Proportion of Subjects Giving Fear of Success Responses in Each Sex–School–Cue Condition

School/sex	Ann cue		John cue	
	Proportion	%	Proportion	%
Coed high school				
Females	9/22	40.9	4/17	23.5
Males	12/39	30.8	2/23	8.7
Noncoed high school				
Females	6/38	15.8	2/34	5.9
Males	9/33	27.3	4/34	11.8

TABLE 2

Proportion of Female Subjects Who Showed Fear of Success in Response to the Ann Cue Classified by Types of Schools Attended

Elementary school	High school			
	Coed		Noncoed	
	Proportion	%	Proportion	%
Noncoed	0/9	0	1/17	5.8
Coed	9/13	69.2	5/21	23.8

Further analyses (summarized in Table 2) revealed that in all but one case in which females responded negatively to the Ann cue, the subjects had attended a coed elementary school ($\chi^2 = 10.25$, $df = 1$, $p < .005$).[4] Of the nine girls who attended a noncoed elementary school and a coed high school, none showed fear of success. Turning to the 34 girls who attended coed elementary schools, we find that the proportion showing fear of success was affected significantly by the kind of high school they attended (69.2% for the coed high school, 23.8% for the noncoed high school; $\chi^2 = 5.09$, $df = 1$, $p < .025$). It might also be

[4] One of the authors visited a kindergarten classroom in one of the coed elementary schools and found, in addition to distinct play areas for boys (blocks, planes, etc.) and girls (a model kitchen), two posters proclaiming (in pink) "what little girls are made of" (sugar and spice...) and (in blue) "what little boys are made of" (snails and puppy dog tails).

noted that the figure of 69.2% is quite similar to the 65% obtained by Horner, most of whose subjects—students at the University of Michigan—had probably always attended coed schools.

Finally, Table 3 shows the frequency with which negative responses to both Ann and John cues fell into affiliative or nonaffiliative categories as a function of sex of subject and type of school. It is evident that negative responses to the Ann cue from both male and female subjects fell primarily into the affiliative category (69.4%), while the reverse pattern obtained for the John cue (only 16.7% in the affiliative category). A 2×2 (Ann–John/affiliative–nonaffiliative) chi-square test indicates that the overall pattern is highly reliable ($\chi^2 = 8.15$, $df = 1, p < .005$).

TABLE 3

NUMBER OF SUBJECTS WITHIN EACH SEX–SCHOOL–CUE CONDITION WHO GAVE AFFILIATIVE OR NONAFFILIATIVE FEAR OF SUCCESS RESPONSES

School/sex	Ann cue		John cue	
	Affiliative	Nonaffiliative	Affiliative	Nonaffiliative
Coed high school				
Females	6	3	1	3
Males	7	5	0	2
Noncoed high school				
Females	6	0	1	1
Males	6	3	0	4

DISCUSSION

The results are compatible with Horner's notion that increased cross-sex competition increases fear of success in females. It appears, however, that the motive is largely formed before a girl reaches high school and that females who have not learned to avoid success in the presence of male classmates in elementary school do not "fear success." For those who do attend coed elementary schools, fear of success is more common, and it appears to be increased by attendance at a coed high school.

Obviously we must be cautious in generalizing from these results. The subject population may not be representative of all high school students. The measure of success avoidance was quite primitive; only after the study was completed did a more precise measuring system involving different cues and different scoring procedures become available (Horner, Tresemer, Berens, & Watson, 1973). Also, we cannot be completely certain that the results are due solely to school rather than home environment. We tried to rule out the effect of home by chosing schools that were as similar as possible while differing on the coed–noncoed dimension. The two high schools are only four blocks apart; the male and female sections of the noncoed school are located in adjacent buildings (although there is no contact between males and females during the school day); the students at the two schools come from similar religious backgrounds, neighborhoods, social classes, and, in several cases, from the same families. In discussions with the principals of the two schools, we were unable to determine any systematic differences between families who chose to send their children to the different schools.

Our results have both theoretical and practical implications. They indicate that developmental studies of fear of success should include preadolescent age groups and should treat school environment as an important variable. On the practical side, the results seem to argue for noncoed schools. Since there may be other arguments that favor coeducation, and considering that most schools in the country would not be likely to change their commitment to coeducation even if the present results proved reliable, it would be more realistic to try to design ways to reduce the motivational handicap developed by females under current arrangements. Some of the relevant issues, for example, the elimination of stereotypes commonly presented in textbooks and in physical education courses, have already been addressed by concerned students and their parents. These are undoubtedly only a beginning.

REFERENCES

Atkinson, J. W. (Ed.) *Motives in fantasy, action, and society*. Princeton, N.J.: Van Nostrand, 1958.

Carnegie Commission on Higher Education. *Op-*

portunities for women in higher education. New York: McGraw-Hill, 1973.

French, E. G., & Lesser, G. S. Some characteristics of the achievement motive in women. *Journal of Abnormal and Social Psychology,* 1964, **68,** 119–128.

Horner, M. S. Sex differences in achievement motivation and performance in competitive and non-competitive situations. Unpublished doctoral dissertation, University of Michigan, 1968.

Horner, M. S. Femininity and successful achievement: A basic inconsistency. In J. Bardwick, E. M. Douvan, M. S. Horner, & D. Gutmann (Eds.), *Feminine personality and conflict.* Belmont, Calif.: Brooks-Cole, 1970.

Horner, M. S. Toward an understanding of achievement-related conflicts in women. *Journal of Social Issues,* 1972, **28,** 157–175.

Horner, M. S., Tresemer, D. W., Berens, A. E,. & Watson, R. I., Jr. *Scoring manual for an empirically derived scoring system for motive to avoid success.* Unpublished monograph, Harvard University, 1973.

Lesser, G. S., Krawitz, R., & Packard, R. Experimental arousal of achievement motivation in adolescent girls. *Journal of Abnormal and Social Psychology,* 1963, **66,** 59–66.

Monohan, L., Kuhn, D., & Shaver, P. Intrapsychic versus cultural explanations of the "fear of success" motive. *Journal of Personality and Social Psychology,* 1974, **29,** 60–64.

Tresemer, D. Fear of success: Popular, but unproven. In C. Tavris (Ed.), *The female experience.* Del Mar, Calif.: CRM Publishers, 1973.

Veroff, J., Wilcox, S., & Atkinson, J. W. The achievement motive in high school and college-age women. *Journal of Abnormal and Social Psychology,* 1953, **48,** 103–119.

(Received November 1, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 731–735

RELATIONSHIPS AMONG PERCEPTUAL–MOTOR TASKS: TRACING AND COPYING[1]

MARCIA BERNBAUM, JACQUELINE GOODNOW,[2] AND ELYSE LEHMAN[3]

George Washington University

Subjects of several ages were asked to copy, trace, or point to the beginning of a set of shapes. The aim was to determine the extent to which sequential behavior (starting points and stroke progressions) was consistent across tasks. With one notable exception, behavior was consistent. The exception was the extent to which children started at the left rather than the right, a behavior that was far less frequent when tracing or pointing to the beginning of a shape than when copying. The results point to the feasibility of using sequential analyses to explore questions of generality and transfer on graphic tasks.

When children copy geometric shapes, they display marked sequential patterns. They start, for example, at the top more often than the bottom, at the left more often than the right, and with a vertical rather than a horizontal stroke (Gesell & Ames, 1946; Goodnow & Levine, 1973; Hanfmann, 1933). Such patterns appear before children learn to write and may help account for the occurrence of particular errors in copying or writing. Shapes like "d," for instance, are reversed more often than shapes like "b" (Goodnow & Levine, 1973; Lewis & Lewis, 1965), an asymmetry of error that appears to stem from children's preferences for starting with a vertical stroke and progressing toward the right.

The sources of such directional behavior are not clear. They might lie, for instance, in some factor specific to the copying task, to handedness, or to the nature of the English script. The present study examines the effect of the task, asking whether sequential behavior is the same on three tasks: copying, tracing, and pointing to the beginning of a shape.

As a variable, the selection of task effects is based on indications that neither handedness nor the nature of the English script offers a sufficient explanation for the directional behavior. Left-handed and right-handed children appear to show similar patterns, the main difference being the extent to which figures are started at the left (Gesell & Ames, 1946). Children in the United States and Israel also display broadly similar patterns, the most prominent differences being in the direction for drawing circles and the age at which most children cease copying shapes with a single, continuous stroke (Goodnow, Friedman, Bernbaum, & Lehman, 1973).

Among tasks, the use of pointing allows a check on the effect of using a graphic tool. The selection of tracing is based on the frequent use of tracing as a technique for improving performance in copying or writing, a usage dating back to the early Greeks and Romans (Fernald, 1943, pp. 27–29). Evidence for the effectiveness of tracing, however, is difficult to find. Most reports of

[1] We are happy to acknowledge a variety of sources of support and assistance. Caroline Eldred, Rochelle Levine, and Marcia Minichiello assisted in the gathering of data. Several schools made work with children possible and pleasant: Beauvoir in Washington, D.C.; Cardinal Forest in Virginia: Bowie Montessori, Carmody Hills, Spring Hill Lake, and Alexander schools in Maryland. Financial support was provided by National Institute of Mental Health fellowship 54542-01 to the senior author and by National Institute of Child Health and Human Development Grants K03-HD-36791 and HD-03105 to the second author.

[2] Now at School of Education, Macquarie University, Sydney 2113, Australia.

[3] Requests for reprints should be sent to Elyse B. Lehman, Department of Psychology, George Washington University, 2125 G St., N.W., Washington, D.C. 20006.

positive effect (e.g. Montessori, 1966) are not experimentally based. Within the small set of experimental studies, Brittain (1969), Goodson (1967) and Rand (1971) report no benefit, while Bee and Walker (1968) report negative transfer. This lack of fit between classroom practice and experimental results suggests at least that the relationship between the two tasks needs additional exploration. It might well be, for example, that children do not spontaneously use the same directional patterns when tracing as when copying. In such a case, the lack of transfer might stem from different patterns being practiced on the two tasks.

A pilot study supported the possibility of a difference between performance when tracing and when copying. One group of children aged 5 to 7 years (n = 60) was asked to copy two shapes (a triangle and a diamond). A group in the same age range, from a different school (n = 60), was asked to trace the same shapes, going over the dittoed outlines with a darker colour. The two groups did not differ in the extent to which they started at the top or in the extent to which they used a single continuous line. They did differ, markedly, however, in the extent to which the first stroke was down the left rather than the right side of each shape. On the triangle, for instance, 84% of the children first drew down the left side when copying. Only 40% did so when tracing, a difference significant, on chi-square analysis, beyond the .01 level. A similar difference occurred with the diamond shape. The use of two different schools, however, left questions about the comparability of the subjects. In addition, treating 5- to 7-year-olds as a single group (a treatment based on one school being ungraded) left questions about effects from changes in age or experience. The present study was designed to cover these questions.

Method

Subjects

For Condition 1 (copying versus tracing), 76 subjects were drawn from four age groups. One group (n = 20) was undergraduate; the other three were children attending a summer day camp in Maryland after completing either kindergarten (n = 20), Grade 1 (n = 16), or Grade 2 (n = 20). Almost all of the children attended, during the school year, the private school operating the day camp. Median chronological age (CA) in the four groups was 20, 6:2, 7:1, and 7:11 years, respectively.

For Condition 2 (copying versus pointing to the beginning of a shape), 52 subjects were drawn from three grades in a public school within Fairfax County, Virginia. Children were in mid-year kindergarten (n = 19), Grade 2 (n = 18), or Grade 4 (n = 16). Median CA was 5:10, 7:8, and 9:9 years, respectively.

For both conditions, all subjects spontaneously used the right hand.

Stimulus Materials and Procedure

Three shapes provided a comparison of tasks: triangle, inverted V, and U-shape. The subjects first copied the three shapes, embedded in a set of 15. Following a distractor task, subjects either traced over the test shapes or pointed to where they thought the shapes began.

Scoring

Productions were first scored for correctness. The liberal criteria are given in Goodnow and Levine (1973). Correct copies were then scored for the observance of three sequential principles. One of these dealt with the overall sequence of strokes, asking whether the sequence showed a thread or a broken line. In a threaded pattern, the subjects lay out a shape as if they were laying out a thread. They may momentarily lift their pencils but no jump occurs in the path. (Scoring for this principle applied only to Condition 1.) The other two principles dealt with starting points, asking whether or not subjects started at a topmost point, and whether they started at a leftmost point or with a stroke to the left. (Scoring for these principles applied to both conditions.)

For each shape, subjects were next scored for whether they displayed same or different behavior across tasks, for example, started at the left or not on both, started at the top or not on both, threaded or not on both. In Condition 2, these counts were reduced to a single score, namely the use of same versus different starting points on both tasks.

The first type of scoring (observance of principles) provides a picture of what subjects did on each task. The second type (presence of same–different behavior) allows a statistical analysis of similarity or consistency across tasks.

Results

Between copying and tracing, a large difference appeared in the incidence of left starting (Figure 1) but not in the incidence of other sequential features (Table 1). The difference was strongest with the kindergarteners. It diminished with age but was still present to some extent among children

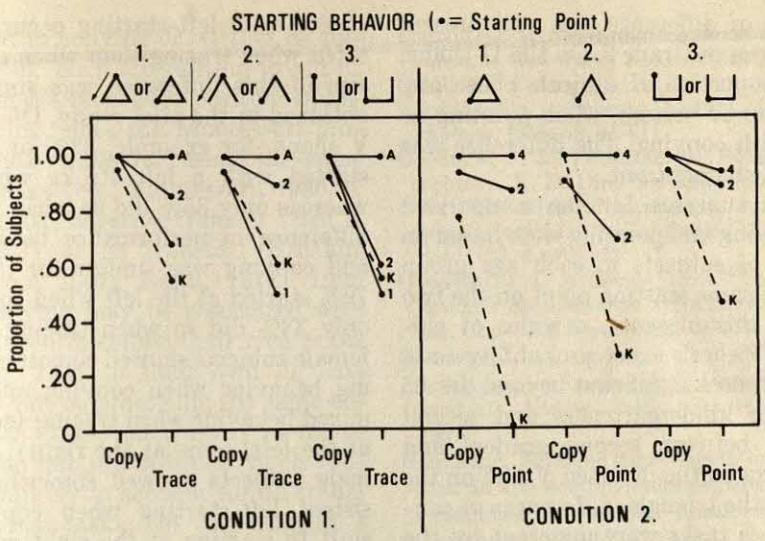


FIGURE 1. The incidence of left-starting behaviors when copying and tracing (Condition 1) or when copying and pointing to the beginning of a shape (Condition 2).

who had completed Grade 2 (Figure 1). Statistically, the three groups of school children were not significantly different from one another in the incidence of same–different behavior across tasks, the only consistent differences being between kindergarteners and adults. For these two groups and for left-starting behavior, Fisher's exact probability tests yielded values of $p = 0.006$ for the triangle, $p = 0.0042$ for the inverted V, and $p = 0.0016$ for the U-shape.

Between copying and pointing, the incidence of starting at the top of the triangle showed variation among the two older groups (Table 1). The most marked difference across tasks, however, again centered around the incidence of left-starting behavior (Figure 1). On the triangle and on the inverted V, a smaller proportion of subjects chose the left base-point when pointing as compared with copying. The effect was strongest in kindergarten: No kindergarteners, for example, pointed to the left base-point as the beginning of the triangle, 53% choosing the top and 47% choosing the right base-point. On the inverted V, a weak form

TABLE I

PROPORTION OF SUBJECTS IN FOUR AGE GROUPS EXHIBITING TWO ASPECTS OF DIRECTIONAL BEHAVIOR

Grade	*n*	Starting at top						Threading					
		Shape 1		Shape 2		Shape 3		Shape 1		Shape 2		Shape 3	
		Copy	Trace	Copy	Trace	Copy	Trace	Copy	Trace	Copy	Trace	Copy	Trace
Kindergarten	20	.50	.70	.35	.35	1.00	1.00	.55	.60	.60	.65	.95	.75
Grade 1	16	.66	.86	.44	.62	1.00	.94	.56	.18	.56	.36	.94	.36
Grade 2	20	.90	.75	.70	.65	1.00	1.00	.15	.25	.30	.35	.80	.35
Adult	20	.75	.55	.25	.10	1.00	.95	.20	.45	.75	.80	.80	.60
		Copy	Point	Copy	Point	Copy	Point						
Kindergarten	19	.47	.53	.45	.39	.74	.89						
Grade 2	18	.33	.55	.39	.11	1.00	1.00						
Grade 4	16	.47	.67	.19	.31	1.00	.94						

Note. Shape 1 = triangle; Shape 2 = inverted V; Shape 3 = U-shape.

of this type of difference was still present among children in Grade 2. On the U-shape, a smaller proportion of subjects chose any left point, top or bottom, when pointing as compared with copying. The difference was restricted to kindergarten.

Statistical analyses for the comparison between copying and pointing were based on the number of subjects in each age group who used the same starting point on the two tasks. With this measure, a series of chi-squares and Fisher's exact probability tests yielded differences significant beyond the .05 level between kindergarteners and second graders and between second graders and fourth graders on the inverted V and on the U-shape. On the triangle, differences in consistency across tasks were undercut by the shift, in the two older groups, toward more top starting when pointing than when copying.

In sum, children after kindergarten and Grade 1 appear to be still moving toward consistency across tasks.

Replication

Although sizable, the difference between copying and tracing was less than that obtained in the pilot study. Such variation might result from a somewhat unstable phenomenon or from the fact that the subjects in Condition 1 had already completed a year in kindergarten and had been taught to write. To explore these possibilities, the comparison of copying with tracing was repeated with an additional group. The children in this group were living in an area where few children attend preschool and where county policy actively discourages formal training, including training in writing, in kindergarten. For the 60 subjects, testing was in the middle of the school year. Median CA was 5:5 years. The shapes were an inverted V and a U-shape. Procedure was the same as in Condition 1.

This comparison established the stability of the difference between tracing and copying. As in Condition 1, no significant differences appeared in the incidence of threading or of starting at the top. A significant difference again occurred, however, in the incidence of starting at the left or with a left stroke, with left-starting occurring far less often when tracing than when copying. The size of this difference was similar to that obtained in the pilot study. On the inverted V shape, for example, 83% of the subjects started with a left stroke when copying, whereas only 35% did so when tracing. The difference in performance between tracing and copying was similar for the U-shape: 79% started at the left when copying, while only 33% did so when tracing. In general, female subjects showed consistent left-starting behavior when copying and shifted to mixed behavior when tracing (some starting at the left, some at the right). In contrast, male subjects showed somewhat less consistent left-starting when copying and a shift to starting at the right when tracing. On the U-shape, for example, 100% of the female subjects started at the left when copying as opposed to 47% when tracing. Males, on the other hand, started at the left 56% of the time when copying and only 19% of the time when tracing.

Discussion

The results suggest that most aspects of directional behavior in copying are not specific to the copying task. The tendency to start at the top rather than at the bottom, for example, was common to all three tasks: copying, tracing, and pointing to the beginning of a shape. In addition, copying and tracing did not differ in the extent to which they elicited the pattern of using a single continuous line. The prominent area of difference lay in the extent to which children displayed starting at the left on all three tasks, this pattern being markedly less frequent when tracing or pointing to the beginning of a shape than when copying. Such variations suggest that the different aspects of directionality may have somewhat different sources. At the least, left–right directionality appears to be the least stable aspect of directionality or the slowest to achieve constant use.

In addition, the results suggest a feasible way of looking at transfer across tasks. One condition for positive transfer may be the use of the same path on both tasks, that is, the insistence by a teacher that the pattern

to be followed in copying or writing be followed in tracing. Such a condition would fit with Montessori's reports of positive transfer from tracing, since the Montessori procedure explicitly calls for using the same path on all tasks (e.g., Montessori, 1966, p. 266). In contrast, negative transfer (Bee & Walker, (1968) or no transfer (Brittain, 1969; Goodson, 1967; Rand, 1971) from tracing to copying may be associated with allowing children to follow their own paths when tracing. Left to their own devices, children may not be practicing the same behavior on both tasks, particularly as far as starting at the left is concerned. Such conditions are unlikely to account for all transfer effects across tasks, but they suggest a minimal safeguard for teachers to adopt when using the technique of tracing and a way of introducing some order into a literature marked by discrepancies.

REFERENCES

Bee, H. L., & Walker, R. S. Experimental modification of the lag between perceiving and performing. *Psychonomic Science*, 1968, **11**, 127–128.

Brittain, W. L. Some exploratory studies in the art of preschool children. *Studies in Art Education*, 1969, **10**, 14–24.

Fernald, G. M. *Remedial techniques in basic school subjects*. New York: McGraw-Hill, 1943.

Gesell, A., & Ames, L. B. The development of directionality in drawing. *Journal of Genetic Psychology*, 1946, **68**, 45–61.

Goodnow, J. J., Friedman, S., Bernbaum, M., & Lehman, E. B. Direction and sequence in copying: The effect of learning to write in English and Hebrew. *Journal of Cross-Cultural Psychology*, 1973, **4**, 263–282.

Goodnow, J. J., & Levine, R. A. "The grammar of action": Sequence and syntax in children's copying. *Cognitive Psychology*, 1973, **4**, 82–98.

Goodson, J. An experimental modification of the perceiving-performing lag. Unpublished undergraduate honors thesis, Stanford University, 1967.

Hanfmann, E. Some experiments on spatial position as a factor in children's perception and reproduction of simple forms. *Psychologische Forschung*, 1933, **17**, 319–329.

Lewis, E. R., & Lewis, H. P. An analysis of errors in the formation of manuscript letters by first grade children. *American Education Research Journal*, 1965, **2**, 25–35.

Montessori, M. The discovery of the child. Madras, India: Kalakshetra Publications, 1966.

Rand, C. S. W. The development of copying skills in drawing: A perceptual versus cognitive representational approach. Unpublished doctoral dissertation, Stanford University, 1971.

(Received November 5, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 736-740

ACQUISITION AND NONSPECIFIC TRANSFER EFFECTS IN PROSE LEARNING AS A FUNCTION OF QUESTION FORM[1]

JAMES F. VOSS[2]

University of Pittsburgh

Acquisition and nonspecific transfer effects in prose learning were studied by presenting a different 2,000-word prose passage on each of three successive days, with each day consisting of four reading and testing trials. Testing employed either multiple-choice or completion questions. Performance failed to improve over passages, indicating an absence of nonspecific transfer effects. Performance on the latter trials tended to deteriorate as a function of passage in the multiple-choice condition. The percentage errors after the first correct response were also greater in the multiple-choice than in the completion condition. The findings provided no evidence that learning-to-learn effects exist with prose materials and suggested that repeated use of the multiple-choice test produces interference at the higher levels of learning.

The study of nonspecific transfer includes the effect of warm-up and the so-called learning-to-learn effect (Postman, 1969). Typically, the study of learning-to-learn has involved the use of traditional verbal learning tasks such as paired-associate learning (e.g., Keppel & Postman, 1966), serial learning (e.g., Meyer & Miles, 1953) and, more recently, free recall (e.g., Postman, Burns, & Hasher, 1970). When prose passages have been used, however, data pertinent to demonstrating nonspecific transfer effects usually have been generated only incidentally, with the evidence suggesting that nonspecific transfer effects, when obtained, are of small magnitude (e.g., Crouse & Idstein, 1972).

The present study was designed to determine whether nonspecific transfer effects could be demonstrated with successive acquisition of samples of unrelated prose materials. The question of whether one would expect nonspecific transfer effects to occur in prose learning may be answered positively or negatively. On the one hand, it could be argued that the process of learning from prose materials is probably quite similar to that found in other learning tasks; therefore, one would expect to find nonspecific transfer effects in the prose context. On the other hand, it could be argued that learning from given samples of prose material will depend not upon training in the task situation but upon factors such as the person's knowledge of the topic and his general level of comprehension skills (Carroll, 1972).

In addition to providing information regarding the possible occurrence of nonspecific transfer effects in prose learning, the present experiment was also designed to study the effects of question form upon acquisition and to determine whether the question form variable differentially influences nonspecific transfer effects. Multiple-choice and completion questions were em-

[1] The research reported herein was supported by the Learning Research and Development Center, supported in part by funds from the National Institute of Education (NIE), U. S. Department of Health, Education, and Welfare. The opinions expressed do not necessarily reflect the position or policy of NIE and no official endorsement should be inferred. The author wishes to thank Cathy Lynch for her assistance in the gathering and tabulation of the data.

[2] Requests for reprints should be sent to James F. Voss, Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

ployed, with the identical stem of the question used in both cases. Theoretically, the multiple-choice and completion questions may be considered as recognition and recall procedures, respectively. Assuming that a recall procedure involves a retrieval process and retrieval in a recognition situation is absent or at least reduced compared to the recall task (e.g., Underwood, 1972), it was expected that completion questions might produce the development of retrieval skills in the given task situation. This would lead to nonspecific transfer effects, while the utilization of such skills would not be found, at least to the same extent, in a multiple-choice recognition task.

Method

Materials

Three prose passages from the World Book Encyclopedia Yearbook (1968) were edited and titled. The three passages pertained to the cultural and geographical development of Formosa, the development of a river project in South Vietnam, and a discussion of Tikal, a Mayan city. The length of each passage was approximately 2,000 words. The passages were unrelated with respect to specific content.

Design and Procedure

Each subject was presented a passage for eight minutes, followed by a seven-minute test trial. Each passage was presented for four such study-test trials. The questions were presented in four different random orders over the four trials. The three passages were presented on three consecutive days, with the order of presentation of the three passages counterbalanced via the use of a randomly permuted 3 × 3 latin square. Question form was a between-subjects variable. The instructions stated that the subjects were to learn what they could from the prose passage in the time allotted and that they would be tested immediately after they finished reading. The nature of the test was not indicated.

Questions

For each of the three passages, 21 questions were constructed in either a completion or a multiple-choice form, with the two forms having identical stems. From preliminary testing involving 24 subjects, the 10 best questions were selected, with selection based upon the consistency with which the question was consistently rated as factual in nature and upon the moderate difficulty level of the question.

Subjects

There was a total of 30 subjects, with 15 subjects serving in the multiple-choice and 15 in the completion condition. The subjects were assigned to the two conditions and to the three passage orders via a table of random numbers. The subjects were introductory psychology students.

Results

The mean correct responses for the three successive passages were 6.70, 7.07, and 6.83, respectively. Analysis of variance revealed that these differences are not significant ($F < 1.00$, $df = 2/56$), thus indicating the absence of significant, nonspecific transfer effects. Furthermore, the Passage × Question Form interaction is not significant ($F < 1.00$, $df = 2/56$).

With respect to the question form variable, multiple-choice questions yielded significantly more correct responses ($M = 7.44$) than completion questions ($M = 6.29$, $F = 6.66$, $df = 1/28$, $p < .05$). In addition, the Question Form × Trials interaction was significant ($F = 4.38$, $df = 3/84$, $p < .01$), as was the Passage × Question Form × Trials interaction ($F = 3.58$, $df = 6/168$, $p < .01$).

The data pertinent to these two significant interactions are shown in Figure 1. The significant three-factor interaction is apparently due to a progressive change over the passages of the Trials × Question Form interaction. Specifically, acquisition of first-passage information increased in virtually a parallel manner for the multiple-choice and completion questions; however, convergence of performance occurred for the two types of questions over trials for the second and third passages. It is of particular interest that on the third and fourth trials of the second and third passages, it was found that multiple-choice performance did not show improvement on the fourth trial. Indeed, inspection of the data of the three passages indicates that performance on the final trial in the multiple-choice condition decreased as a function of passage. Thus, the data suggest that completion questions produced consistent improvement over trials for all three passages, although there is no learning-to-learn effect; however, mul-

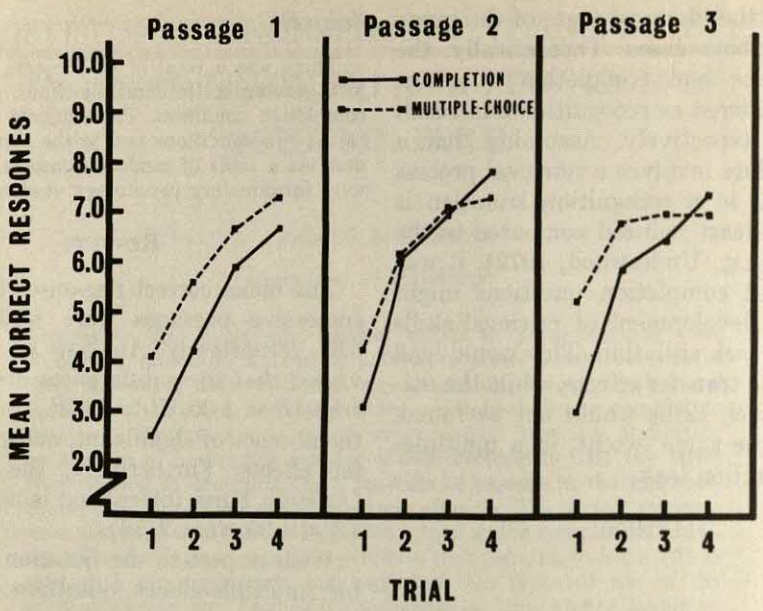


FIGURE 1. Mean correct responses as a function of trial for the completion and multiple-choice question forms for each successively presented passage.

tiple-choice questions produce consistent improvement only on the first passage while on the second and third passages, multiple-choice performance deteriorates at the higher levels of acquisition.

The Figure 1 data also suggest the possibility that Trial 1 performance may be a factor in the Passage × Trial × Question Form interaction, since the data at least suggest that significant positive transfer may have occurred on Trial 1 as a function of passage, especially for the multiple-choice condition. However, analysis of the Trial 1 data indicated that although multiple-choice Trial 1 performance was significantly superior to that found with completion questions ($F = 22.18$, $df = 1/29$, $p < .01$)—a result that is not surprising—there nevertheless was not a significant effect of Passage ($F = 1.20$, $df = 2/56$, $p > .05$) or of the Passage × Question Form interaction ($F < 1.00$, $df = 2/56$).

A more detailed analysis of the completion and multiple-choice conditions provided the following information. First, as one would expect, the completion condition produced a number of omission errors, namely, .48, .41, and .45 of the total number of errors for the three respective passages. Second, the percentage errors after the first correct responses were tabulated by determining the number of opportunities for error after the first correct response for each passage for each subject, dividing the number of opportunities into the number of errors, and multiplying by 100. The three respective passages of the multiple-choice condition yielded a percentage error after the first correct response of 7.3%, 4.9%, and 8.5%, respectively; for the completion condition, the respective percentages were .4%, 3.2%, and 3.5%. Analysis revealed a significant difference of question form condition, ($F = 5.57$, $df = 1/28$, $p < .05$), thus indicating that the multiple-choice procedure involved some difficulty in maintaining stability of correct response. Tabulation of the errors made after the first correct response indicated that over the three passages, errors were made on 10 different responses in the completion condition and 42 different questions in the multiple-choice condition. Moreover, if 2 errors after the first correct response occurred on the same question, the probability was .76 that the same error would be repeated in the multiple-choice

condition. There were 17 cases of 2 or more errors after the first correct response in the multiple-choice condition and only 5 in the completion condition. Of these 5, 3 involved repetition of the same error. Thus, in the multiple-choice condition, there is considerable difficulty in maintaining correct responses, and when errors are made and repeated, they tend to be the same errors.

Discussion

The failure to obtain nonspecific transfer effects in the present study is probably attributable to one of two factors. One possibility is that the present experiment may be insensitive to nonspecific transfer effects, although it is difficult to see in what way the experiment may be viewed as deficient in this regard. The most probable reason is that at the level of the college student, prose comprehension may be "content-determined" and not "task-determined," that is, for experienced readers, the understanding of prose may primarily depend upon how well the individual comprehends the specific content of the material; for less experienced readers, however, learning from prose may be much more related to the level of development of comprehension skills. From this reasoning, one might expect that nonspecific transfer may occur in children who are in the earlier phases of reading development. It also may be noted that Slamecka (1960) obtained nonspecific transfer effects with connected discourse, using the verbatim recall procedure for 20-word sentences. This result may be due to the fact that with such a procedure, Slamecka's task and measurement procedures were more closely related to the traditional serial-learning paradigm than those of the present study, with Slamecka's design placing strong emphasis upon rote memory and upon minimizing the need to comprehend the contents of the passage.

The second aspect of the present study which warrants comment is the following finding: With repeated passages, the multiple-choice procedure tended to yield less effective performance at higher levels of acquisition with an increase in the number of passages acquired. The locus of this effect appears to be in the tendency for a greater percentage of errors to be made following a correct response in the multiple-choice than in the completion condition. The most likely reason for this result is that the test phase of the multiple-choice method involved exposure to incorrect responses, and these responses may have subsequently interfered with stabilizing the correct responses. One implication of this interpretation is that when one uses multiple-choice questions to test how well a given passage has been learned, subsequent testing may produce errors because of interference attributable to the incorrect alternatives. The multiple-choice method, in other words, may not simply measure performance but may be detrimental to long-term retention. In effect, the multiple-choice procedure of the present experiment may be viewed as a modified retroactive inhibition paradigm in which the stem represents A and the possible responses as indicative of Associations A–B, A–C, A–D, and A–E. On four successive presentations, there is a tendency for the subject to store the incorrect responses as well as the correct responses. When subsequent recall is required, the subject has trouble discriminating correct from incorrect responses, especially at the higher levels of acquisition. Because these responses are the last to be learned, they would, virtually by definition, involve the most difficult questions. This interpretation is in general agreement with the findings of Myrow and Anderson (1972), who attributed the occurrence of retroactive inhibition in the multiple-choice, prose-learning situation to competition of responses of original and interpolated learning. The present study, even though not constituting a basic retroactive inhibition paradigm, extends this finding in that the results suggest the testing procedure itself may be involved in producing interference, and such interference could quite conceivably be reflected in subsequent testing.

REFERENCES

Carroll, J. B. Defining language comprehension: Some speculations. In J. B. Carroll & R. O.

Freedle (Eds.), *Language comprehension and the acquisition of knowledge*. Washington, D. C.: Winston, 1972.

Crouse, J. H., & Idstein, P. Effects of encoding cues on prose learning. *Journal of Educational Psychology*, 1972, **63**, 309–313.

Keppel. G., & Postman, L. Studies of learning to learn: III. Conditions of improvement in the performance of successive transfer tasks. *Journal of Verbal Learning and Verbal Behavior*, 1966, **5**, 260–267.

Meyer, D. R., & Miles, R. C. Intralist-interlist relations in verbal learning. *Journal of Experimental Psychology*, 1953, **45**, 109–115.

Myrow, D. L., & Anderson, R. D. Retroactive inhibition of prose as a function of the type of test. *Journal of Educational Psychology*, 1972, **63**, 303–308.

Postman, L. Experimental analysis of learning to learn. In G. H. Bower & J. T. Spence (Eds.), *The psychology of learning and motivation*. New York: Academic, 1969.

Postman, L., Burns, S., & Hasher, L. Studies of learning to learn: X. Nonspecific transfer effects in free recall learning. *Journal of Verbal Learning and Verbal Behavior*, 1970, **9**, 707–715.

Slamecka, N. J. Retroactive inhibition of connected discourse as a function of similarity of topic. *Journal of Experimental Psychology*, 1960, **60**, 245–249.

Underwood, B. J. Are we overloading memory. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory*. Washington, D. C.: Winston, 1972.

World Book Encyclopedia Yearbook. Chicago: Field Enterprises Educational Corporation, 1968.

(Received November 7, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 741–749

STUDENT TEAMS AND LEARNING GAMES: THEIR EFFECTS ON CROSS-RACE AND CROSS-SEX INTERACTION[1]

DAVID L. DEVRIES[2] AND KEITH J. EDWARDS[3]

Center for Social Organization of Schools, Johns Hopkins University

Although many public schools are nominally desegregated, the interaction among students of varying racial and ethnic backgrounds is minimal. The present study evaluated the effects of two teaching techniques—student teams and instructional games—on the level of cross-race and cross-sex interaction in the classroom. Placing students on heterogeneous four-member student teams created significantly greater cross-race and cross-sex helping and friendship. Team success did not have the predicted positive effect on cross-race and cross-sex interaction. Playing the instructional game had a marginal effect on cross-race helping only; however, the game–team combination considerably increased the incidence of cross-race and cross-sex interaction over that of games alone. Katz's theory of biracial work groups was evaluated in light of the present results.

Even as many school systems are desegregating by altering the racial composition of their schools, social integration of minority groups remains minimal. Integrated education demands far more than changing the racial composition of schools and classrooms; it demands establishing working relationships and understanding among the students and between teachers and students (Georgeoff, Jones, Bahlke, & Howard, 1970; Wagoner, Glatt, & Gaines, 1970; Winnecoff & Kelly, 1971). Desegregated schools need to restructure the classroom in order to create more positive and constructive relationships among students from varying backgrounds. In the present study, the independent and combined effects of student teams and instructional games on both cross-race and cross-sex interaction in the classroom are examined.

There is considerable evidence that barriers to racial interaction exist among students in desegregated schools. Mann (1959), Webster (1960), St. John (1964), Gottlieb (1965), and Cusick (1973) found considerable within-race preference for both classmates and friends by both black and white students. Such within-race preference has been substantiated for students from the sixth grade through college. As noted by both McPartland (1968) and Pettigrew (1969), the racial barriers are stronger for the more intimate relationships, such as friendship, than for formal task relationships.

The existence of barriers to interaction among students across sex lines, particularly for children in early adolescence, has been frequently noted (Coleman, 1961;

[1] This report is part of a program of studies examining structural variations in the organization of both classrooms and schools, and was sponsored by funds from the U.S. Office of Education and National Institute of Education, U.S. Department of Health, Education, and Welfare.

The authors are indebted to the following individuals: Layman Allen, the creator of the instructional game used, for his patient and invaluable assistance in the design of the experimental treatment; Robert Freeman, Paul Gorman, Terry Lazarus, George McCulloh, Joyce Morrow, and Laurie Schaefer, staff members of Northern Parkway Junior High School, Baltimore, Maryland, for opening up their school and classrooms; Gail Fennessey, Alyce Nafziger, and John Snyder for their assistance in the collection of data; John Holland, Samuel Livingston, and James McPartland for their critical reading of earlier drafts; Carol DeVries for her valuable editorial comments on the several drafts, and Ann F. Forthuber for typing the manuscript.

[2] Requests for reprints should be sent to David DeVries, Center for Social Organization of Schools, Johns Hopkins University, Baltimore, Maryland 21218.

[3] Now with Rosemead Graduate School of Psychology, Rosemead, California.

Waller, 1967). Abel and Sahinkaya (1962) found strong within-sex preference even among four- and five-year-olds. Cross-sex relationships among junior high school students are particularly uncommon and, if they do occur, are often strained. Although barriers to cross-sex interaction have been commonly noted for early adolescents, the interpersonal dynamics of such barriers have not been explored. Because the dynamics of racial barriers are more clearly delineated and at first glance appear to be similar to those for cross-sex interaction, the authors used the racial prejudice literature for predicting structural effects on both cross-race and cross-sex interaction.

Recent reviews of the race relations literature (cf. Amir, 1969; Pettigrew, 1969) suggest reasons why merely creating desegregated classrooms is not sufficient for improving race relations among the students. One important condition for effective cross-race contact appears to be the creation of multiracial interdependencies among the representatives of the various racial or ethnic groups. One way to structure such interdependencies is the creation of biracial student work groups (or teams) in which similar rewards are administered to all teammates. The use of biracial student teams has frequently been suggested as a way to create greater interracial cooperation and acceptance (Allport, 1954; Gottlieb, 1965; Thelen, 1970).

A variety of empirical studies support the use of group or team rewards as a means of improving race relations (cf. Amir, 1969; Katz, 1970). However, as noted by Katz (1970), the ex post facto nature of the majority of the studies reviewed leaves unclear the causal direction of the team reward/attitude change relationship. An examination of the experimental studies of biracial work teams reveals mixed results. Witte (1972) created biracial student teams in college classrooms and observed their effects on interracial acceptance and interaction. The biracial team condition consisted of students performing frequently on individual tasks. However, teammates' scores were combined to form a team score. A team grade was calculated and assigned to all teammates. Team practice sessions, during which peer-tutoring activities among teammates were allowed, constituted the only in-class cross-racial contact. The primary cross-racial bond was created by the reward interdependency. When compared to control condition scores (students performing individually and being rewarded on the basis of individual performance), the biracial team treatment created greater interracial peer tutoring, more racial acceptance (measured by several racial attitude scales), and less racial isolation (measured by a seating aggregation index). Witte's results can be interpreted as supporting Allport's (1954) hypothesis that interracial acceptance can be increased if a common goal is created among teammates by making them cooperatively dependent upon each other.

Katz (1970) describes two experimental studies (Katz & Benjamin, 1960; Katz, Goldston, & Benjamin, 1958) which also examined the effects of forming cooperative dependence across racial lines on interracial acceptance. In both studies, subjects were placed in four-member work groups (consisting of two black males and two white males). Subjects in all conditions were assigned a variety of tasks with a group product required for the vast majority of the tasks. Cooperative dependence was experimentally manipulated by varying the level of reward from the individual to the group. In the *group-reward* condition, subjects received a bonus payment based on the performance of the group as a whole. In the *individual-reward* condition, each subject received a bonus payment based on his performance alone. For both studies, the comparison between the group- and individual-reward conditions revealed no systematic differences in interracial acceptance (as reflected by communication patterns).

The apparent contradiction between the results of Katz and his associates (Katz & Benjamin, 1960; Katz et al., 1958) and those of Witte (1972) is explainable. An examination of the treatment manipulation in the studies of Katz suggests that a high level of cooperative dependence was created even in the individual-reward condition. Although students were rewarded individually, the vast majority of the tasks were structured so as to require considerable collaboration

among the subjects. That is, resources required to complete the task were shared by the group as a whole which created a mutual dependence, even though rewards were ultimately based on individual performance. In contrast, the control condition in the Witte (1972) study created no such interdependence among students. Because the studies conducted by Katz confounded task with reward interdependence, they constitute weak tests of the effects of team reward on race relations.

Because the present study uses groups similar to those of Witte and groups which meet most of Allport's conditions for effective racial interaction, the authors predict increased cross-race interaction when students are placed on biracial work teams. The authors postulate a similar effect on cross-sex interaction for students placed on teams containing both boys and girls. In addition, Amir (1969) has noted that for biracial work groups to be effective in reducing cross-racial barriers, the group must be operating under high-reward contingencies. That is, biracial groups that are highly successful at the task assigned to them should evidence greater dissolution of racial barriers than groups which fail at the task. This hypothesis is tested for both cross-race and cross-sex interaction in the present study.

No studies have dealt with the effects of playing games on cross-race and cross-sex interaction and acceptance. A number of empirical studies (cf. Boocock & Schild, 1968) reveal a positive effect of game playing on student attitudes and academic achievement. However, playing a learning game places students of different races and/or sexes in face-to-face competition, and both Allport's (1954) model and the studies by Katz (1970) suggest that interracial contact in competitive settings may be aversive to the individuals. In short, games do force students of different races and of opposite sex to interact on a pleasant task; however, the interaction is competitive. Therefore, although the predicted effect of games on cross-race and cross-sex interaction is positive, it is less than the predicted effect of teams because of the competitive nature of the interaction in a game setting.

Method

Subjects

The subjects were 110 seventh-grade students at a junior high school located in a major east-coast metropolitan area; 43% of the students were blacks and 47% were males. The students came from predominantly lower-middle-class families.

Design

The experiment was a 2 × 2 randomized design; the two factors were task (game versus quiz) and reward (team versus individual). Students were randomly assigned to the four treatment conditions. The experiment was conducted over a 4-week period, and it involved 20 school days. All groups met during the first period of the day. Two male and two female teachers participated in the study. All were in their first or second year of teaching. At the midpoint of the study, the four teachers were rotated; consequently, each treatment was taught by both a female and a male teacher. There was no significant nonrandom clustering by race or sex of students across treatment conditions. The racial distribution ranged from 38% to 52% blacks, and sex distribution ranged from 41% to 50% males.

Independent Variables

The experiment manipulated two major variables—academic task and reward structure. The academic task placed students in either a quiz or game-playing performance situation. The reward structure involved the administering of rewards to either individual students or teams of students.

The weekly schedule of all treatment groups proceeded as follows: Students performed on the math task (game or quiz) for half of the period each Tuesday and for the entire period on Friday. Practice sessions were conducted for half of the period each Tuesday and Thursday; these sessions were separate from the game and quiz activities. For the remainder of the week, the students received traditional classroom instruction.

Academic task. One task level consisted of semiweekly teacher-made math quizzes on material covered during the preceding days. For each quiz, a student was assigned a percentage score based on the number of problems answered correctly. The second level consisted of semiweekly game sessions with students performing on the instructional game, *Equations* (Allen, 1969). Within a classroom of 30 students, 10 *Equations* games were played simultaneously (3 students per game). The players at each table were grouped homogeneously on mathematics achievement. At the end of each game session, the 3 students at a table were ranked according to their performance. The winner received 6 points, the middle scorer 4 points, and the low scorer 2 points. Students who were high scorers were "bumped" up to a higher achievement table

for the next game-playing session. Low scorers were assigned to a lower achievement table for the following session. Such bumping maintained homogeneous game tables while taking into account new learning (as reflected in a student winning or losing).

Reward structure. Both reward structures used a classroom newsletter as a means of providing public reinforcement for performance on either the quiz or game. The newsletters were handed out to all students in each class on the day following every game session or quiz. The *individual-reward* condition involved the ranking of scores (based on either game or quiz performance) of individual students on the newsletter. Students who scored particularly high or who evidenced substantial improvement in their scores were singled out for special attention. For the *team-reward* condition, the newsletter listed the preceding day's scores for each team (first page) and for each student (second page). The team scores were ranked, and particular praise was given to the top teams or teams moving up rapidly in the rankings. The student teams consisted of four members and were designed with the purpose of creating maximal intrateam heterogeneity (on race, sex, and achievement) and interteam equality. Teammates were assigned adjacent seats and were allowed to work together during practice sessions. However, each team member did perform individually in both the game and quiz classes. The team score listed on the newsletter was obtained by summing the scores of all members present for the game or test. Teams with absent members were thus penalized, and no provisions were made for make-ups.

At no point during the study were the teachers or students made aware of the cross-race or cross-sex hypothesis nor was such interaction overtly encouraged in any of the groups.

Dependent Variables

Three sociometric items were included in a posttest questionnaire in order to assess the level and direction of student helping and friendship patterns. The students were asked to give the names of classmates (*a*) "whom you helped," (*b*) "who helped you," and (*c*) "who are your friends." Eight blank lines were allotted for responses to each of the three questions.

An examination of the interstudent agreement for the friendship choices revealed 61% mutual choices. A similar examination of the you helped/helped you items revealed 46% agreement. The slight reduction in interstudent agreement on the helping relationship (as compared to friendship) is due to less agreement within the game classes. Students in the quiz classes agree about helping relationships almost twice as often (62%) as those in the game classes (33%). The difference is statistically significant ($\chi^2 = 28.08$, $df = 1$, $p < .001$). The nature of the student interaction in the game condition may have made it difficult for students to assess helping patterns.

Hypotheses

Of interest in the present study are the effects of the task and reward factors on the number of cross-race and cross-sex selections (over all selections) on both helping and friendship dimensions. It is hypothesized that both the game and team factors increase cross-race and cross-sex interaction, with the team effect being the larger of the two.

The effect of level of reward on cross-race and cross-sex interaction within the two team conditions was also of interest. Clearly, in the classroom setting of this experiment it was neither desirable nor feasible to manipulate the level of rewards received by the team. Level of reward was made contingent on team performance. Consequently, the low- and high-reward conditions were determined in a post hoc manner. The two teams in each group which won most consistently were classified as high reward, and the two teams which were consistently last were classified as low reward. While the authors recognize that this procedure departs from the approach typically used in experimental social psychology, it is more consistent with behavior in actual classrooms.

Results

The significance of the task, reward, and interaction effects were tested using the log-linear model for hypothesis testing in multidimensional contingency tables (Goodman, 1970). The log-linear model for contingency tables is analogous to the linear model for analysis of variance. The main and interactive effects are defined as linear functions of the logarithms of cell frequencies in the same manner as such effects are defined by linear functions of cell means in analysis of variance. When the total N of the contingency table is large, the variance of the effects is estimated by the sum of the reciprocals of cell frequencies. The ratio of the effect to the square root of its variance (referred to in the present paper as a Z ratio) follows an asymptotic normal distribution and is the statistic used to test the significance of the effects. One-tailed tests of significance were employed for variables analyzed with the log-linear model (see Goodman, 1969, 1970; Shaffer, 1973, for more detailed discussions of the model).

The dependent variables of interest are the percentages of cross-race and cross-sex interaction for helping, being helped, and friendship. Analyses were conducted separately for the Race × Sex interaction dimensions.

Cross-Race Interaction

The log-linear model was used to examine the effects of task, reward, and the Task × Reward interaction on the level of cross-race interaction. A significant task effect was noted for the "helped you" variable only ($Z = 1.68$, $p < .05$). Significant reward effects were observed for "you helped" ($Z = 1.70$, $p < .05$) and "helped you" ($Z = 2.41$, $p < .01$), and a marginally significant effect was detected for friendship ($Z = 1.60$, $p < .10$). Teams resulted in greater cross-race interaction for all three sociometric measures. No significant Task × Reward interaction effects were obtained.

The treatment cell percentages for cross-race helping and friendship are contained in Table 1. The level of cross-race selection for the individual-quiz treatment can be treated as a measure of interracial interaction in a traditional class. The table indicates 33% cross-racial selection for "you helped," 20% for "helped you," and 31% for "friends." A definite preference by students for friends and helpmates of the same race existed in the traditional class. The significant, positive team effect can be noted by comparing the percentages of the top two rows with those in the bottom two rows.

The effect of level of team reward on cross-race interaction was examined by comparing the percentages of cross-race selections for the high-reward teams to those of the low-reward teams. Level of reward was defined by success in the tournament. The chi-square test for difference in proportions was used (Ferguson, 1966). None of the three chi-squares computed was statistically significant ($p < .05$). An examination of the cell percentages reveals a trend in a direction opposite of that predicted (you helped: low reward = 55%, high reward = 39%; helped you: low reward = 53%, high reward = 35%; friends: low reward = 44%, high reward = 33%); that is, low-success teams appeared to create more, not less, cross-race selections on all three dependent variables.

Cross-Sex Interaction

The log-linear model applied to the cross-sex selection indicated no significant task effects. Significant reward effects were noted for all three sociometric variables (you helped: $Z = 2.61$, $p < .01$; helped you: $Z = 3.09$, $p < .01$; friends: $Z = 2.88$, $p < .01$). A significant Task × Reward interaction was observed for the "helped you" variable only ($Z = 2.29$, $p < .05$). Table 2 contains the percentage of cross-sex selections for the four treatment conditions. The traditional class (individual quiz) has minimal cross-sex interaction (you helped = 13%, helped you = 10%, friends = 21%). The significant, positive team effect can be detected by comparing the percentages in the top two rows with those in the bottom two rows.

Figure 1 delineates the significant interaction effect noted for the "helped you" dimension. It appears that adding games to the curriculum had a positive effect on the level of cross-sex helping for the individual-reward condition and a slightly negative effect for the team-reward condition.

The effect of level of team success on cross-sex selection was also assessed and re-

TABLE 1

PERCENTAGE OF CROSS-RACE SELECTIONS FOR THE FOUR TREATMENT CONDITIONS

Treatment	You helped	Helped you	Friends
Individual			
Quiz	33 (15)	20 (20)	31 (131)
Game	20 (30)	29 (49)	27 (108)
Team			
Quiz	38 (60)	34 (61)	37 (159)
Game	44 (52)	54 (56)	34 (147)

Note. Numbers in parentheses indicate total number of student selections.

TABLE 2

PERCENTAGE OF CROSS-SEX SELECTIONS FOR THE FOUR TREATMENT CONDITIONS

Treatment	You helped	Helped you	Friends
Individual			
Quiz	13 (15)	10 (20)	21 (131)
Game	27 (30)	33 (49)	17 (108)
Team			
Quiz	43 (60)	49 (61)	33 (159)
Game	46 (52)	41 (56)	27 (147)

Note. Numbers in parentheses indicate total number of student selections.

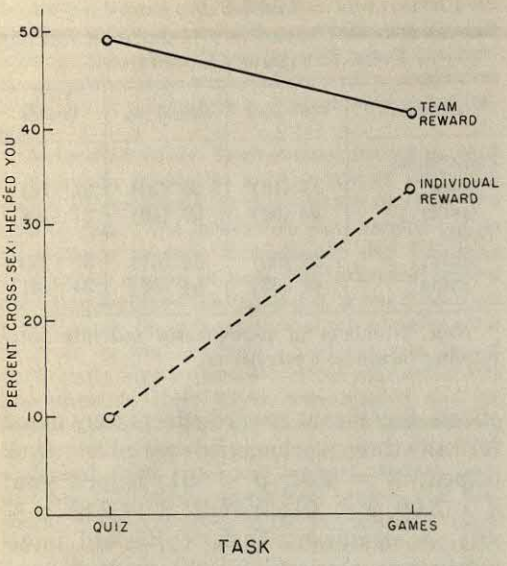


FIGURE 1. The Teams × Games interaction for cross-sex selections.

vealed a marginally significant effect for the "you helped" variable only ($\chi^2 = 2.93$, $df = 1, p < .10$). The cell percentages were as follows: You helped: low reward = 52%, high reward = 30%; helped you: low reward = 47%, high reward = 45%; friends: low reward = 35%, high reward = 27%. The effect on cross-sex interaction for the "you helped" variable was in the opposite direction of that predicted.

Summary of Results

The results indicate a positive impact of team rewards on cross-sex selections in the classroom for both helping and friendship interactions. A positive team effect was also noted for cross-race selections, although the effect was limited to helping or on-task interactions. The instructional game, particularly when used without student teams, had a much less pronounced effect than did teams. For cross-sex selections the game had no effects, and for cross-race selections a positive effect was noted only for the "helped you" dimension. Team success did not influence the level of either cross-sex or cross-race selection.

DISCUSSION

The results indicate that administering team rewards to heterogeneous (on both race and sex) groups of students helps reduce race and sex barriers inhibiting student interaction. The results partially support those of Witte (1972) in which group rewards increased interracial interaction and acceptance. The team effect on reducing cross-racial barriers was particularly potent for the helping relationships. Only minimal generalization of the team effect to friendship dimensions was detected, calling into question whether or not basic changes in racial attitudes were created. For the cross-sex dimension, the positive effect of group reward on cross-sex helping behavior generalized to the friendship dimension.

The success of the team rewards in reducing barriers to social interaction may be due to the following factors. First, the group composition was uniquely heterogeneous. The group composition of the vast majority of the teams consisted of two blacks (one male and one female) and two whites (one male and one female). Perhaps forming such a heterogeneous group of students in the classroom, placing the students in adjacent seats, and allowing them to work together (without administering any group reward) would have had an effect on race relations similar to that observed in the present study. It would have been useful to implement an additional condition involving heterogeneous groups (but with individual rewards) in the current study. As it now stands, it is unclear whether the team-reward effect is due to the team reward or to some combination of the team-reward and group-composition factors.

A second factor in explaining the positive team-reward effect lies in the likely impact of the mutual interdependence (among teammates) on group processes. The reward interdependence created at the team level fosters interaction among teammates. Because the teammates share control over each other's fate in the classroom, it is to their advantage to collaborate with one another and to foster each other's progress in the class. In an analysis of additional data from the present study (DeVries & Edwards, 1973), observations of team practice sessions revealed that substantial amounts of within-team peer tutoring occurred in the team-reward classes. Students in the team classes

also reported greater mutual concern among their classmates than did those in the individual-reward condition (DeVries & Edwards, 1973). In short, team rewards appeared to alter the nature and extent of interpersonal relationships among the students. Because the teams formed in the present study were uniquely heterogeneous, such extended interpersonal relationships occurred across racial and sex lines.

An important distinction in the operationalization of student teams in the present study as compared to those of Katz et al. (1958) and others should be noted. Traditionally, biracial student teams involved teammates working on a group task. For example, a group of students might be asked to solve a specific human relations problem, arriving at a group solution. The group's solution would then be rated, with each teammate receiving the group grade. In the current study, each teammate, acting as a representative of his team, performed individually on either the quiz or game. Within-team interaction was not demanded, although time was allotted for semiweekly practice sessions during which teammates could interact. The use of individual performance settings, although justifiable on a pedagogical basis (it forces a higher level of accountability of student performance), probably had the effect of diluting the team-reward effect on cross-race and cross-sex interaction. Subsequent research by the authors will investigate the effect of team rewards on reducing social barriers using (*a*) individual performance by teammates and (*b*) group performance by the entire team.

Amir's (1969) hypothesis that strong rewards must be administered to the team if racial barriers are to be dissolved was not supported. Although the evidence is only suggestive, teams which consistently lost in the competition appeared to have slightly greater cross-race and cross-sex helping and friendship relationships than those which won. Perhaps failure at a task by a team forces the individuals to realize the limits of their own resources and the importance of cooperating with their teammates. In contrast, individuals on teams which are consistently successful are not confronted with the need for help from others and thus are not forced to break down racial and sex barriers.

Unfortunately, because of the post hoc classification of the teams in the present study, the observed results might also be due to a different causal chain. More specifically, teams which attempted greater cross-race and cross-sex interaction may have used up considerable time and energy on such activities, leaving little time for actual preparation for the tournament. In contrast, teams which did not attempt to dissolve the barriers may have rather spent their time in individual study preparing for the tournament; that is, forming social relationships within a team may be a costly process which detracts from the academic performance of the teams. In short, although Amir's hypothesis was not supported, more rigorous tests of the reward hypothesis are required before a definitive answer can be achieved.

In seeking an explanation for the lack of a significant game effect on reducing race and sex barriers, one possible reason might lie in the interaction defined by the games. It is possible that few of the three-person groups at the game tables were actually biracial. An examination of the racial composition at the game tables over the entire experimental period revealed that, on the average, 78% (range from 71% to 84%) of the students were playing at racially mixed tables on a given day. In short, it appears that the game treatment did create many biracial interaction situations but that such situations did not foster greater interracial friendships or helping relationships.

The lack of a significant game main effect is possibly due to the student competition created by the game. Although students in the *Equations* classes encounter students of the opposite sex and different races at the game tables, they do so in the context of meeting an "opponent." The outcome of such an encounter is always a ranking of the individuals on some status dimension. Students playing the game are not likely to go out of their way to assist their fellow players (their opponents). Additionally, the games–individual treatment involves small game-table groups whose membership constantly changes. Even if two students de-

sired to become friends, the two would soon be separated because each day they face different opponents. Because *Equations* (without teams) creates competition among students and because the composition of the groups changes daily, it fails to overcome race and sex barriers in the classroom. However, when teams are added to the game structure (as advocated by its designer, Layman Allen), the combination facilitates both cross-race and cross-sex interaction.

Work by Katz (1970) on biracial groups indicates it may be important to examine the team effects for black and white students separately. Katz views biracial groups as combining a high-status group (whites) and a low-status group (blacks). Given this theory, one would predict different reactions from whites and blacks to being on a biracial team. Whites as the high-status group might be threatened by the inclusion of blacks, particularly when included in their friendship groups. An examination of the cross-race data for the team-condition students revealed no significant differences (use of chi-square test) between the two racial groups in percentage of cross-racial selection for the three sociometric variables (you helped: blacks = 49%, whites = 35%; helped you: blacks = 38%, whites = 48%; friends: blacks = 44%, whites = 33%). The data suggest that the biracial interaction in the teams classes was not one of a low-status group trying desperately to gain entry into a high-status group (as suggested by Katz, 1970) but rather of two equal-status groups mutually sharing needed resources.

The current study suggests that relations between racial and sex groups can be improved in the classroom if the teacher restructures the reward system. Administering group rewards (to heterogeneous student groups) represents a classroom technique readily accessible to teachers across both grade levels and subject matter. Recent studies of team rewards in classrooms (Hamblin, Buckholdt, Ferritor, Kozloff, & Blackwell, 1971; DeVries & Edwards, 1973; Witte, 1972; Wodarski, Hamblin, Buckholdt, & Ferritor, 1971) suggest that such rewards create high levels of student peer tutoring and, in turn, increase general academic achievement. This study, in addition, suggests such tutoring can take place between students of different races and of opposite sexes, thus contributing to the academic achievement of the students involved and advancing the integration process in desegregated schools.

The study leaves unanswered several questions concerning the generalization of the observed effects on social interaction of minority groups in the classroom. First, there is the issue of whether the observed experimental effects on overt behavior (using self-report measures) generalized to the affective level. The marginal effect of team rewards on cross-race friendships suggests little attitudinal change. In contrast, the cross-sex interaction increased for both the formal task-related dimensions (helping) and the more intimate social dimensions (friendships). Whether such generalization failed to occur even for cross-sex interaction could only be answered by more direct attitudinal measures. A second unanswered question concerns the generalization of the treatment effects over time. Although it is unlikely that the increased cross-race and cross-sex interaction would persist once the team-reward structures were dropped, it is at this point an open question subject to future investigation.

A question raised earlier in the paper was whether structural variations which have been posited as reducing natural barriers to interaction across racial lines would act similarily on barriers across sex lines. The results of the present study suggest that they do. Placing students on teams characterized by a diversity of personal characteristics of teammates, a common team goal, and equal-status positions for all teammates not only increases cross-race interaction but has even greater positive effect on cross-sex interaction. Both natural barriers (race and sex) appear to operate under similar dynamics.

One possible limitation of the current results is the unique racial and sex distributions within all classes of the present experiment. All classes were characterized by an almost 50–50 split along both race and

sex lines. If, for example, only 20% of the students in a class were one race or one sex, such groups would truly represent a minority in the classroom. In such classrooms, the heterogeneous student teams might not create a major reduction in racial or sex barriers as observed in the present study. Answers to questions such as these await further research.

REFERENCES

Abel, H., & Sahinkaya, R. Emergence of sex and race friendship preferences. *Child Development*, 1962, **33**, 939–943.

Allen, L. E. *Equations—A game of creative mathematics*. Wff N' Proof, Turtle Creek, Pa. 1969.

Allport, G. *The nature of prejudice*. Cambridge, Mass.: Addison-Wesley, 1954.

Amir, Y. Contact hypothesis of ethnic relations. *Psychological Bulletin*, 1969, **71**, 319–342.

Boocock, S. S., & Schild, E. O. (Eds.) *Simulation games in learning*. Beverly Hills, Calif.: Sage, 1968.

Coleman, J. S. *The adolescent society*. New York: Free Press of Glencoe, 1961.

Cusick, P. A. Black-white student interactions in an urban secondary school. Paper presented at the annual convention of the American Educational Research Association, New Orleans, February 1973.

DeVries, D. L., & Edwards, K. J. Learning games and student teams: Their effects on classroom process. *American Educational Research Journal*, 1973, **10**, 307–318.

Ferguson, G. A. *Statistical analysis in psychology and education*. New York: McGraw-Hill, 1966.

Georgeoff, P. J., Jones, I., Bahlke, S. J., & Howard, O. The effect of the curriculum on racial cleavage in integrated fourth grade classrooms. *Contemporary Education*, 1970, **41**, 297–303.

Goodman, L. A. How to ransack social mobility tables and other kinds of cross classification tables. *American Journal of Sociology*, 1969, **75**, 1–40.

Goodman, L. A. Multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, 1970, **65**, 226–256.

Gottlieb, D. School integration and absorption of newcomers. *Integrated Education*, 1965, **3**, 69–75.

Hamblin, R. L., Buckholdt, D., Ferritor, D., Kozloff, M., & Blackwell, L. *The humanization processes: A social, behavioral analysis of children's problems*. New York: Wiley, 1971.

Katz, I. Experimental studies of Negro-white relationships. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. Vol. 5. New York: Academic Press, 1970.

Katz, I., & Benjamin, L. Effects of white authoritarianism in biracial work groups. *Journal of Abnormal and Social Psychology*, 1960, **61**, 448–465.

Katz, I., Goldston, J., & Benjamin, L. Behavior and productivity in biracial work groups. *Human Relations*, 1958, **11**, 123–141.

Mann, J. H. The effect of interracial contact on sociometric choices and perceptions. *Journal of Social Psychology*, 1959, **50**, 143–152.

McPartland, J. The relative influence of school and of classroom desegregation on the academic achievement of ninth grade negro students. *Journal of Social Issues*, 1968, **25**, 93–102.

Pettigrew, T. F. Racially separate or together? *Journal of Social Issues*, 1969, **26**, 43–69.

Shaffer, J. P. Defining and testing hypotheses in multidimensional contingency tables. *Psychological Bulletin*, 1973, **79**, 127–141.

St. John, N. H. De facto segregation and interracial acceptance in high school. *Sociology of Education*, 1964, **37**, 326–344.

Thelen, H. A. A proposal for the attainment of racial integration through public education. *School Review*, 1970, **78**, 391–396.

Wagoner, J. L., Jr., Glatt, C. A., & Gaines, W. A. Student relationships and activities in desegregated schools. *High School Journal*, 1970, **54**, 188–201.

Waller, W. *The sociology of teaching*. New York: Wiley, 1967.

Webster, S. W. The influence of interracial contact on social acceptance in a newly integrated school. *Journal of Educational Psychology*, 1960, **51**, 292–296.

Winnecoff, H. L., & Kelly, E. W., Jr. Problems in school desegregation: Real or imaginary. *Integrated Education*, 1971, **9**, 3–10.

Witte, P. H. The effects of group reward structure on interracial acceptance, peer tutoring, and academic performance. Unpublished dissertation, Washington University, 1972.

Wodarski, J. S., Hamblin, R. L., Buckholdt, D., & Ferritor, D. Effects of individual and group contingencies on arithmetic performance. Paper presented at the annual meeting of the American Educational Research Association, New York, January 1971.

(Received October 29, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 750–755

EFFECT OF INTENT-ACTION DISCREPANCY AND STUDENT PERFORMANCE FEEDBACK ON TEACHER BEHAVIOR CHANGE[1]

WALTER DOYLE[2]
University of Notre Dame

JUDITH McNALLY REDWINE
Indiana University at South Bend

The researchers studied the influence of two types of verbal feedback on changes in teacher perceptions and behavior. The experiment was conducted in a microteaching-type laboratory setting in which 36 experienced junior high school teachers taught the same content to two groups. Feedback between sessions consisted of information about (*a*) discrepancy between stated intent (subject's estimate of how he or she would teach) and observed behavior and (*b*) student learning outcomes. To isolate feedback effects, prompts concerning desired behavior and other training dimensions were removed from treatments. Teacher behavior was coded using interaction analysis categories. Analysis of variance indicated that feedback treatments were assocated with significant changes in teacher intent ($p < .05$). No significant effects were observed for teacher behavior or for the interaction of treatments. Results suggest that knowledge of student learning outcomes and the reduction of intent-action discrepancy have minimal impact on teacher decisions to modify classroom procedures.

The development and use of systematic observation procedures in teacher training has focused attention on feedback processes and their effects on teacher performance. A cursory glance at recent literature indicates that feedback experiences have become a standard component of proposals for teacher education reform (e.g., Smith, 1969). Research on feedback effects, however, has often lacked adequate experimental controls and has been limited to a narrow range of feedback content. Moreover, the processes assumed to mediate teacher change from feedback have seldom been examined directly.

The present study was designed to investigate, under controlled conditions, the effects of feedback on the perceptions and behavior of experienced teachers. In contrast to most studies in this area, content was held constant across teaching trials, and feedback concerning both classroom behavior processes (perceived and observed) and student learning outcomes was provided between sessions. Finally, research questions were formulated in terms of discrepancy reduction, a process frequently assumed to operate in teacher reaction to feedback.

Research in the interaction analysis tradition (e.g., Amidon & Hough, 1967; Flanders, 1970) has generally been interpreted as indicating that coded feedback concerning classroom processes is an effective means for modifying teacher behavior. Flanders (1963, 1964, 1965) has suggested that these feedback effects can be explained in terms of teacher reduction of intent-action discrepancy. That is, interaction analysis feedback, offered in an objective and nonthreatening manner, allows a teacher "to compare his actions with his intentions" on the assumption that "any reduction of the discrepancy between what a teacher intends to do and what he actually does is likely to result in an improvement of instruction [Flanders, 1965, p. 13]."

[1] The authors wish to thank R. H. Metzcus for assistance in the conduct of the research and J. Walter for help in preparing the report.

[2] Requests for reprints should be sent to Walter Doyle, who is now at the College of Education, North Texas State University, Denton, Texas 76203.

Unfortunately, existing interaction analysis research does not permit firm conclusions about the effects of feedback or of intent-action discrepancy reduction. Studies in the Flanders' tradition have typically employed "training in interaction analysis," a treatment which, in addition to feedback, includes discrimination training, reinforcement, practice, and strong prompting in favor of "indirect" teaching behaviors. It is difficult, therefore, to separate the effects of feedback and intent-action discrepancy reduction from the effects of other treatment components, especially the learning of an approved set of teaching behaviors.

In one of the few studies of discrepancy reduction in teacher behavior change, Tuckman, McCall, and Hyman (1969) made an effort to remove bias toward indirect behavior from interaction analysis training materials and to control for the magnitude of initial discrepancy between self-perceived and observed teaching behavior. They found, however, that high-discrepancy and low-discrepancy teachers did not change behavior differentially. Rather, discrepancy reduction was associated with a modification of self-perception toward greater congruence with observed performance. These results suggest that, in the absence of an external model, reduction of discrepancy induced by process feedback results in a modification of personal intentions rather than teaching behavior.

Surprisingly, feedback research has concentrated almost exclusively on the effects of process content. It would seem, however, that a teacher's decision to change behavior operates within a matrix of information inputs, of which process feedback represents only one dimension. Presumably a teacher evaluates a procedure also in terms of its "success," that is, its impact on student learning outcomes. Such an evaluation would seem to influence a decision to modify behavior. Moreover, if a teacher is confronted with information about student learning outcomes *and* classroom processes, the nature and degree of change would seem to be a function of possible interactions between these two sources or types of feedback.

From a discrepancy reduction viewpoint, studies of teacher attribution, a process whereby persons assign causal responsibility for events and outcomes in an environment (Heider, 1958), have implications for formulating research questions about the effects of student performance feedback. In simulated teaching situations, Johnson, Feigenbaum, and Weiby (1964) and Beckman (1970) found that teachers accepted personal credit for student success (attribution to self) and assigned responsibility to outside factors for student failure (external attribution). In these studies, no attempt was made to relate attribution of causality for student performance to changes in perceived or observed teaching behavior. The results suggest, however, that insofar as student failure induces a discrepancy between intentions and outcomes, teachers resolve this discrepancy in a manner which does not necessarily involve behavior change. That is, if a teacher perceives himself responsible only for those methods that are successful, then there is little need to modify his performance when confronted with student failure.

The purpose of the present study was to examine the effects of verbal feedback concerning (*a*) student learning outcomes and (*b*) the discrepancy between self-reported intent and observed action. Consistent with previous research on discrepancy reduction, effects of feedback were measured in relation to teacher changes in both perceptions of how they would teach (intent) and actual classroom performance as measured by independent observers. Teachers receiving feedback that their students were not successful were expected to change behavior to a greater extent than teachers who were successful. However, differential changes resulting from interaction effects were predicted in two instances: (*a*) Teachers told that their actions did not match their intentions but that their students were successful were expected to modify intent; and (*b*) teachers told that their actions did not match their intentions and that their students were not successful were expected to change behavior.

Method

Subjects

The sample consisted of 36 junior high school teachers who volunteered to participate in an experiment on microteaching. Subjects included 12 female and 24 male teachers representing language arts, math, science, social studies, art, health, and occupational education. Subjects averaged 7.25 years of teaching experience.

Procedure

The experiment was conducted in a microteaching-type laboratory setting in which subjects taught two 10-minute sessions to different groups of four students selected, within practical limits, on a random schedule. A 3 × 2 factorial design permitted the manipulation of intent-action discrepancy (high–low) and student performance (success–nonsuccess–control) feedback conditions.

Teaching content consisted of six Greek prefixes and suffixes and their English meaning (e.g., *tele*—far). On the day before they were to teach, subjects were introduced to the teaching task and were asked to complete an intent inventory. This intent measure was derived from the 10 categories of interaction analysis (Flanders, 1970) and required that each subject estimate the percentage of time he expected to devote to various classroom behaviors during the teaching session.

Following the first teaching session, subjects were told to wait while teaching behavior records were calculated and students were tested on the material just taught. After an appropriate interval (approximately 10 minutes), subjects received feedback regarding intent-action discrepancy and student test performance. Subjects were then given a second teaching task consisting of a different but equivalent set of Greek words and English meanings. Subjects also completed the intent inventory for the second teaching session. An observer was present during both teaching sessions, and audio recordings of each session were made. The teach–feedback–reteach cycle was completed in approximately 40 minutes.

Feedback was administered in both written and oral form by a research assistant unfamiliar with the nature of the experiment. Feedback content consisted first of a summary of observed teaching behavior (calculated in minutes) presented in conjunction with the subject's stated intent as measured prior to the teaching session. Subjects in the success group were also informed that the mean score for their students on the postsession test was "significantly *above* the average as previously established for this test." Subjects in the nonsuccess group were told that their students scored below the average for the test, and control subjects received no student performance feedback. In order to avoid sensitizing subjects to a particular pattern of teaching behavior, no further discussion or elaboration of this information was permitted during the feedback session.

The mode of feedback was designed to correspond closely to the "verbal feedback" treatment used by Tuckman et al. (1969). These researchers found that verbal feedback from an independent observer, in contrast to interaction analysis training (without feedback) and to audiotape playback, was the only treatment associated with a significant change in teacher behavior. This feedback mode was selected for the present study because it appeared to be a potentially effective treatment which was not dependent upon prior familiarity with interaction analysis and which contained minimal external cues concerning "approved" teaching behaviors. Interviews with subjects used in the present research indicated that they had received no prior interaction analysis training or feedback.

Assignment to Treatments

Since the present study focused on the effects of the discrepancy reduction process itself, rather than on the direction of teacher change, all subjects received the same classroom behavior feedback content. The use of standard feedback content also served to equalize treatment conditions in order to control for a possible source of variation. For purposes of the study, classroom behavior feedback content was derived from composite matrix data provided by Amidon and Flanders (1963) and consisted of 85% total talk, 57% total teacher talk, 28% student talk, 15% indirect teacher influence, and 42% direct teacher influence. Subjects were assigned to intent-action discrepancy feedback conditions according to their "discrepancy scores," that is, the differences between intent and the process feedback they were to receive. Discrepancy scores were computed from the absolute value of the difference between stated intent as measured prior to the first teaching session and the standard feedback content selected for the study. Subjects with discrepancy scores above the median were assigned to the high-discrepancy group and those below the median to the low-discrepancy group. Subjects within each discrepancy group were then randomly assigned to the three student performance feedback treatments: success, nonsuccess, and control.[a]

Because of the element of deception involved in both feedback treatments, each subject was informed about the nature of the study immediately following the second teaching session. Responses to inquires during these debriefing sessions indicated that subjects did not doubt the validity of the feedback content they received.

---

[a] In assigning subjects to conditions, discrepancy level was treated as a moderator variable (Tuckman, 1972, pp. 41–43, 110–113). As a result, subjects were not randomly assigned to discrepancy levels, but "within each level on the moderator variable," subjects were randomly assigned to student performance feedback treatments "to insure that all other selection characteristics . . . were randomized across conditions . . ." (Tuckman, 1972, p. 110).

TABLE 1

ANALYSIS OF VARIANCE OF BEHAVIOR SCORE CHANGE ($|B_2 - B_1|$) BY TREATMENTS AND HIGH–LOW DISCREPANCY

Source	*df*	Total talk		Teacher talk						Student talk	
				Total		Direct		Indirect			
		MS	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*
Treatments	2	19.60	1.12	23.09	1.56	50.88	1.37	12.89	.65	10.43	.43
High–low discrepancy	1	2.89	.17	9.51	.64	13.45	.36	9.82	.50	9.00	.37
Interaction	2	25.77	1.47	25.42	1.72	14.79	.40	1.42	.07	17.63	.73
Within	30	17.48		14.77		38.52		19.70		24.17	

Dependent Variables

Dependent variables consisted of two categories: changes in teacher intent and changes in teaching behavior. Changes were scored in five areas: total talk, total teacher talk, student talk, direct influence, and indirect influence. Intent changes were derived by computing the absolute values of the difference between pretreatment intent inventory scores and posttreatment intent scores. Behavior changes were computed from the absolute value of the difference between actual teaching behavior in the two teaching sessions. Actual teaching behavior was coded from audiotapes by two trained observers using the Flanders system of interaction analysis. Mean interobserver agreement over the 72 sessions, using the Scott procedure (see Flanders, 1967), was 81.7.

RESULTS

Change in Teaching Behavior

Table 1 summarizes analysis of variance results indicating the effects of student performance feedback treatments and discrepancy conditions on the five dependent measures of teaching behavior change. There were no statistically significant effects for discrepancy level or for treatments. No significant effects were found for the interaction of treatments and discrepancy conditions.

Change in Teacher Intent

Table 2 summarizes analysis of variance results concerning the effects of student performance feedback treatments and discrepancy conditions on the five measures of teacher intent change. The data indicate that discrepancy conditions had a significant effect on the amount of intent change in the areas of direct influence ($p < .05$) and indirect influence ($p < .05$). A comparison of means for the two groups revealed that high-discrepancy subjects ($\bar{X} = 27.44$ for direct influence and $\bar{X} = 59.78$ for indirect influence) changed more than low-discrepancy subjects ($\bar{X} = 12.61$ for direct and $\bar{X} = 17.50$ for indirect) on these two measures of teacher intent. The data further indicate that there was no significant effect on intent change for student performance feedback treatments and no significant interactions between treatments and discrepancy conditions.

Because of possible selection bias introduced by the method of assigning subjects to treatments (see Footnote 3), chi-square analyses were conducted on the distribution of sex and experience across discrepancy levels. Results for sex were not significant ($\chi^2 = 1.13, df = 1, p > .20$).[4] Results for experience indicate a moderate trend ($\chi^2 = 2.92, df = 1, p < .10$)[4] toward an inverse relation between these two variables, that is, high-experience subjects were assigned to the low-discrepancy level and low-experience subjects to the high-discrepancy level (see Table 3).

DISCUSSION

Under the conditions of the present study, feedback concerning student learning outcomes did not have a significant impact on the modification of teacher perceptions of how they would teach or on actual classroom performance. Moreover, the interaction of student performance and class-

[4] Chi-square calculations were based on Tuckman's (1972, pp. 248–249) formula which incorporates a correction for continuity.

TABLE 2

ANALYSIS OF VARIANCE OF INTENT SCORE CHANGE ($| I_2 - I_1 |$) BY TREATMENTS OF HIGH-LOW DISCREPANCY

Source	*df*	Total talk		Teacher talk						Student talk	
				Total		Direct		Indirect			
		MS	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*
Treatments	2	384.03	.66	289.58	.88	32.86	.12	870.53	.29	75.69	.47
High-low discrepancy	1	1,667.36	2.86	802.78	2.45	1,980.25	7.34*	16,086.69	5.34*	367.36	2.29
Interaction	2	213.18	.37	234.02	.71	354.25	1.31	1,913.50	.63	17.36	.11
Within	30	583.19		327.50		269.62		3,013.78		160.69	

* $p < .05$.

room process feedback did not produce differential changes in either teacher intent or behavior. Resolution of intent-action discrepancy was, however, associated with a modification of teacher intent in the areas of direct and indirect influence. These results suggest that teachers, when asked to reteach the same content, react to intent-action discrepancy and student performance feedback by revising their intentions of how they will teach rather than their classroom behavior, regardless of whether or not these procedures were effective.

Present findings must be interpreted, however, in light of possible limitations inherent in the design. First, the experimental setting, although allowing for control of extraneous variables, may have been sufficiently artificial to minimize potential for behavior change. Consistency with results obtained under more natural conditions (e.g., Tuckman et al., 1969) would seem, however, to support the validity of the present findings. A second limitation concerns possible selection bias introduced by the method of assigning teachers to treatments. The analysis indicates a trend toward differential experience levels across high- and low-discrepancy conditions. Although experience may be associated with receptivity to feedback, existing research has not confirmed such a relationship (Tuckman & Oliver, 1968). Moreover, it is unlikely that selection bias would have reduced potential for behavior change or would have been associated with systematic changes in teacher intent as indicated in the present data. It would seem, therefore, that selection bias is not a serious threat to the utility of the findings.

TABLE 3

DISTRIBUTION OF SUBJECTS BY EXPERIENCE ACROSS DISCREPANCY LEVELS

Discrepancy level	Experience[a]	
	High	Low
High	4	14
Low	10	8

[a] Subjects with years of experience above $\bar{X} = 7.25$ were classified as high experience; those below the mean as low experience.

Within these limitations, results of the present study suggest that the personal intentions of experienced teachers do not serve as performance models of sufficient power to change classroom behavior. In the absence of an externally validated and supported model of approved behavior, verbal feedback would appear to affect perceptions rather than behavior. This interpretation implies that the modification of teacher behavior, in contrast to perceptions, is dependent on contingencies other than those provided by verbal feedback. Contingency analysis, along the lines suggested by McDonald (1973), would seem to represent a profitable direction for clarifying teacher behavior change research.

Implicit in these findings is the notion that knowledge of student learning out-

comes has little impact on teacher decisions to change behavior. Given the fact that student performance feedback is the only type available to teachers on a regular basis, further research on ways to optimize the effect of such feedback would seem to be imperative.

REFERENCES

Amidon, E. J., & Flanders, N. A. *The role of the teacher in the classroom.* Minneapolis, Minn.: Paul S. Amidon, 1963.

Amidon, E. J., & Hough, J. B. (Eds.) *Interaction analysis: Theory, research and application.* Reading, Mass.: Addison-Wesley, 1967.

Beckman, L. Effects of students' performance on teachers' and observers' attributions of causality. *Journal of Educational Psychology,* 1970, **61,** 76–82.

Flanders, N. A. Intent, action and feedback: A preparation for teaching. *Journal of Teacher Education,* 1963, **14,** 251–260.

Flanders, N. A. Some relationships among teacher influence, pupil attitudes, and achievement. In B. J. Biddle & W. Ellena (Eds.), *Contemporary research on teacher effectiveness.* New York: Holt, Rinehart & Winston, 1964.

Flanders, N. A. *Helping teachers change their behavior.* (Rev. ed.) Ann Arbor: University of Michigan, 1965.

Flanders, N. A. The problems of observer training and reliability. In E. J. Amidon & J. B. Hough (Eds.), *Interaction analysis: Theory, research and application.* Reading, Mass.: Addison-Wesley, 1967.

Flanders, N. A. *Analyzing teacher behavior.* Reading, Mass.: Addison-Wesley, 1970.

Heider, F. *The psychology of interpersonal relations.* New York: Wiley, 1958.

Johnson, T. J., Feigenbaum, R., & Weiby, M. Some determinants and consequences of the teacher's perception of causation. *Journal of Educational Psychology,* 1964, **55,** 237–246.

McDonald, F. J. Behavior modification in teacher education. In C. E. Thoresen (Ed.), *Behavior modification in education.* The seventy-second yearbook of the National Society for the Study of Education. Part 1. Chicago: University of Chicago Press, 1973.

Smith, B. O. *Teachers for the real world.* Washington, D.C.: American Association of Colleges for Teacher Education, 1969.

Tuckman, B. W. *Conducting educational research.* New York: Harcourt Brace Jovanovich, 1972.

Tuckman, B. W., McCall, K. M., & Hyman, R. T. The modification of teacher behavior: Effects of dissonance and coded feedback. *American Educational Research Journal,* 1969, **6,** 607–619.

Tuckman, B. W., & Oliver, W. F. Effectiveness of feedback to teachers as a function of source. *Journal of Educational Psychology,* 1968, **59,** 297–301.

(Received November 19, 1973)

Journal of Educational Psychology
1974, Vol. 66 No. 5, 756–771

EFFECTS OF A FOUR-YEAR FOLLOW THROUGH PROGRAM ON ECONOMICALLY DISADVANTAGED CHILDREN[1]

WILLA D. ABELSON AND EDWARD ZIGLER[2]

Yale University

CHERYL L. DEBLASI

Hamden–New Haven Cooperative Education Center, Connecticut

Longitudinal and cross-sectional comparisons were made of economically disadvantaged children in one Follow Through and several non-Follow-Through primary school programs. A longitudinal comparison was also made of the economically disadvantaged and nondisadvantaged children in the Follow Through program. Clear-cut program effects were found in the fourth year (Grade 3) when Follow Through children were superior to non-Follow-Through children on IQ, achievement, and social-motivational measures. The Follow Through program did not result in the economically disadvantaged children attaining the level of intellectual achievement displayed by the nondisadvantaged children. The findings were discussed in relation to current issues in the field of compensatory education.

Optimism concerning the effectiveness of preschool compensatory programs such as Project Head Start has waned considerably in the last few years. Findings have been generally consistent that at the end of a year of Head-Start-type experience, children are superior to comparison children without this preschool experience in both intellectual and social-emotional functioning (cf. Beller, 1969; Klaus & Gray, 1968; Weikart, 1971; Zigler & Butterfield, 1968). However, with some exceptions (Ryan, 1974), the evidence also indicates that the superiority of Head Start children vanishes or is greatly diminished by the end of one year of formal public school (Bronfenbrenner, 1974; Westinghouse Learning Corporation, 1969).

In the face of this evidence, some have concluded that compensatory education in general, or Head Start in particular, is a failure (Eysenck, 1971; Jensen, 1969). Others (Kohlberg, 1968; Zigler, 1973) have responded to these findings with the view that the expectation of long-term benefits

[1] This study was supported by Research Grant MH-03008 from the National Institutes of Mental Health, U. S. Public Health Service; the Gunnar Dybwad Award of the National Association for Retarded Citizens; and the Follow Through Project of the U. S. Office of Education. The authors wish to express their appreciation to the staffs of the Hamden–New Haven Cooperative Education Center and the Hamden, Hartford, and New Haven Public Schools for their cooperation and assistance.

We are indebted to many individuals for their help during the course of the four-year investigation. In particular, we wish to thank Robert S. Avery, Director of the Hamden–New Haven Cooperative Education Center; Samuel Nash, Director of Special Projects and Planning for New Haven Public Schools; Alexander J. Plante, former Chief, Bureau of Compensatory and Community Education Services, State of Connecticut; Grayce Dowdy and Elizabeth Levine, New Haven Public Schools Prekindergarten Program; Martin Grosse, Indiana State University; and C. Allison, L. Eberle, J. Fischer, N. Godfrey, G. Hymer, E. Katz, B. Levine, H. Matthews, E. Meid, D. Metz, B. Saia, I. Segal, S. Styfco, and W. Travis, research staff.

We are grateful to Stephen Klein and Guy Strickland, Center for Evaluation of Education, University of California, Los Angeles, for permission to use a prepublication form of the Attitude Toward School Questionnaire (ASQ); and to Victoria Seitz and Sally Styfco for their critical reading of the manuscript.

[2] Requests for reprints should be sent to Edward Zigler, Department of Psychology, Yale University, New Haven, Connecticut 06510.

from a one-year preschool program was an unrealistic one, since it was based on the questionable assumption that children could somehow be inoculated against the social, economic, and educational disadvantages which they would encounter later in life. Within this view, compensatory education would be seen as having a positive impact on the lives of economically disadvantaged children provided the effort were of long enough duration. At the level of national social policy, this viewpoint resulted in the creation of the Follow Through project.

The Follow Through project sponsors four-year compensatory education programs for both Head Start graduates and non-Head-Start children. The project is an experimental effort permitting a variety of pedagogical approaches. (See Maccoby & Zellner, 1970, for a description of the ly models employed.) Although Follow Through programs have been in operation for several years, available national evaluation data are limited to longitudinal findings for children who have been in the program for only one year (Stanford Research Institute, 1971). This initial evaluation revealed enough differences in favor of Follow Through children to lead some to conclude that the project is a success, although others regard the results as disappointing in view of the relatively high cost of the project and the small absolute size of the differences between Follow Through and non-Follow-Through children. Officials responsible for the Follow Through project have taken the position that "a definitive interpretation of the first-year findings must await the results of ongoing evaluation efforts." [3]

The major purpose of the present paper is to report the findings of one such evaluation effort. A Follow Through program based on a model which has been employed in a number of Follow Through centers was studied for four years. The effectiveness of the program was assessed through longitudinal data, cross-sectional data, and a special longitudinal comparison of economically disadvantaged and nondisadvantaged Follow Through children.

The longitudinal portion of the evaluation involved 35 economically disadvantaged children who attended the full four years (kindergarten through third grade) of the Follow Through program and 26 comparison children who were attending three non-Follow-Through schools during this same period. In view of the small numbers remaining in the longitudinal study at the end of the fourth year, the decision was made to also conduct a cross-sectional study comparing the performance of the entire group of 42 economically disadvantaged children graduating from the Follow Through program (7 children transferred to the program from other schools during the first and second years) with the entire group of 100 economically disadvantaged children of the same age and grade in four non-Follow-Through schools (74 children plus the 26 longitudinal comparison children).

The 35 longitudinal Follow Through children were also compared with 10 nondisadvantaged children who had attended the full four-year Follow Through program. This final comparison was conducted in order to illuminate two issues: (*a*) the progress of nondisadvantaged children attending school classes composed primarily of economically disadvantaged children and (*b*) the comparative performance of economically disadvantaged and nondisadvantaged children under this special program.

Method

Subjects

Longitudinal. At the outset of the study, the subjects included 61 children attending Follow Through (FT) kindergartens and 48 children attending kindergartens in three non-Follow-Through (NFT) schools. All of the children resided in New Haven, Connecticut, and all were from economically disadvantaged families, defined as follows: (*a*) lived in low-income housing, (*b*) parents had no more than a high school education, and (*c*) parents were employed as semiskilled or unskilled workers or were unemployed. The FT group was recruited for the program by soliciting parents in several low-income areas. The NFT group was made up of all of the economically disadvantaged children in one classroom from each

[3] Wilson, R. C. Communication from the U. S. Department of Health, Education, and Welfare, Office of Education, Washington, D. C., February 13, 1969.

TABLE 1

CHARACTERISTICS OF LONGITUDINAL GROUPS

School program	*n*	*M* CA at school entrance (years, months)	Sex		Head Start		Race		Father absent	
			Boys	Girls	Yes	No	Black	White	Yes	No
Follow Through	35	5–1	21	14	32	3	30	5	15	20
Non-Follow-Through	26	5–2	12	14	7	19	19	7	5	21

Note. Abbreviation: CA = chronological age.

of three schools located in similar low-income areas.

Upon completion of the longitudinal study four years later, 35 of the original 61 FT children and 26 of the original 48 NFT children were completing third grade in these schools. Three of the original FT children and three of the original NFT children were held back in kindergarten or first grade and were dropped from the longitudinal samples. The others who were dropped withdrew from these schools during the course of the study. An examination of the children's demographic characteristics and their test scores at kindergarten entrance revealed no systematic differences between either (*a*) those who stayed in or dropped out of the Follow Through program or (*b*) those who stayed in or dropped out of the non-Follow-Through programs.[4]

The characteristics of the economically disadvantaged longitudinal groups are presented in Table 1. The groups differed in four respects: As compared to the non-Follow-Through group, the Follow Through group had (*a*) a higher proportion of boys than girls, (*b*) a higher proportion of blacks than whites, (*c*) a higher incidence of Head Start attendance, and (*d*) a higher incidence of father absence. In addition to these children, 10 nondisadvantaged children (5 girls, 5 boys; 4 black, 6 white) completed four years of the Follow Through program. These children resided in two-parent homes located in middle-income neighborhoods and the occupations of their fathers ranged from lower-middle to upper-middle class (e.g., fireman, salesman, professor).

Cross-sectional. The cross-sectional sample consisted of all third-grade children in the Follow Through program and in three non-Follow-Through programs who were economically disadvantaged (as defined by the criteria given above), who were completing their fourth year of school, and for whom English was the primary language. The FT cross-sectional group consisted of 42 children: the 35 longitudinal children plus 7 children who had transferred to the FT program during kindergarten or first grade. The NFT group of 100 children consisted of the longitudinal NFT group plus an additional 74 children. This group was made up of the following school samples:

1. Inner-city tutorial ($n = 48$): nine longitudinal children and their 35 classmates in a New Haven inner-city school with a tutorial program, plus 4 longitudinal children in another New Haven inner-city school with the same program;

2. Outer-urban enriched ($n = 30$): thirteen longitudinal children and their 17 economically disadvantaged classmates in a New Haven school attended by both economically disadvantaged and and nondisadvantaged children; and

3. Inner-city traditional ($n = 22$): twenty-two children in an inner-city school 35 miles from New Haven.

The characteristics of the FT and NFT cross-sectional groups are presented in Table 2. As was true of the longitudinal groups, the cross-sectional groups differed in proportion of boys and girls, incidence of Head Start attendance, and incidence of father absence. Unlike the longitudinal groups, the cross-sectional groups had similar proportions of blacks and whites.

School Programs

Follow Through. The Follow Through program under study was conducted by the public schools of New Haven and Hamden, Connecticut. The program was based from its inception on the educational model which is usually identified with Bank Street College of Education (see Maccoby & Zellner, 1970), and staff from Bank Street were involved in the training of Follow Through teachers during the last two years of the study. The FT classes were composed of no more than 20 children and at least two full-time staff (head teacher and assistant teacher). Although the majority of children were economically disadvantaged, each class included some children from middle- to upper-middle-income families.

The features which most markedly distinguished the FT program from programs in the NFT schools in the study were (*a*) an individual rather than a group-oriented approach, (*b*) an explicit interest in the social-emotional development of the child, or what is often called the "whole child" approach, and (*c*) an emphasis on learning how to learn through the mastery of underlying principles and concepts. These features led to more frequent contacts between teachers and individual children

[4] Comparisons involving the larger samples tested in kindergarten and first grade have been reported elsewhere (Abelson, 1974).

TABLE 2

CHARACTERISTICS OF CROSS-SECTIONAL GROUPS TESTED AT END OF THIRD GRADE

School program	*n*	*M* CA (years, months)	Sex		Head Start		Race		Father absent	
			Boys	Girls	Yes	No	Black	White	Yes	No
Follow Through	42	8–9	25	17	37	5	37	5	17	25
Non-Follow-Through	100	8–10	51	49	26	74	89	11	27	73

Note. Abbreviation: CA = chronological age.

in FT classrooms than in NFT classrooms. They also resulted in the use of a broad array of teaching methods, for the manner in which a lesson was conducted in an FT classroom depended on the needs, interests, and ongoing responses of the children in that classroom, rather than on a preset curriculum or technique. The FT program emphasized verbal communication skills. In addition, adult–child relations were oriented specifically toward fostering children's self-esteem and interpersonal trust.

Non-Follow-Through. In the longitudinal study, three NFT schools in New Haven were employed; these same schools plus a school in another Connecticut city were employed in the cross-sectional study. The four schools served neighborhoods which were comparable to those of the economically disadvantaged FT children. Three of the schools (including the one outside New Haven) were in inner-city neighborhoods with predominantly black and Spanish-speaking populations. The other school was in a non-inner-city New Haven area with a 60% black, 40% white population. Children attending this school either came from low-income families residing in a public housing project or from middle-income families in the surrounding neighborhood.

At the beginning of the study, the three schools in New Haven had quite similar, traditional public school programs. Classes were larger than in the FT program, and classroom teachers seldom had more than occasional outside assistance. Lessons were usually organized for each class group as a whole, with the exception of reading which was traditionally taught with small groups.

During the second year of the longitudinal study, class sizes were reduced and a number of experimental projects were initiated in the New Haven schools. In the two inner-city schools attended by NFT children, low-achieving children began to be tutored individually (several NFT children in the present study were tutored). In one of these schools, some children were placed in a busing program to schools in middle-class neighborhoods (five of the original NFT children were selected for this program and were therefore dropped from the study). The non-inner-city school in the socioeconomically mixed neighborhood was not involved in these projects, but the facilities of this school were greatly expanded at this time and the school staff initiated an extensive program of extracurricular activities.

The programs of these three NFT schools thus changed in some significant respects during the course of the longitudinal study. Although the pedagogical approaches used in the classrooms continued to reflect the group-oriented, didactic model which has traditionally been followed in public schools, the range of educational opportunities that were available to the children broadened considerably. It was primarily because of these changes that the fourth NFT school, located in another city, was sought out and added to the cross-sectional study. As the children in this fourth NFT school had not received any special remedial or enrichment programs, they provided a sample of comparison children with a more traditional type of inner-city school experience than the children in the other NFT schools.

Measures

Academic achievement. Academic achievement at the beginning and end of kindergarten was measured with the Screening Test of Academic Readiness (STAR; Ahr, 1966). The STAR is a group-administered instrument designed to appraise general information, conceptual maturity, and perceptual-motor development in preschool- and kindergarten-aged children. Academic achievement in Grade 1 was assessed with the Metropolitan Achievement Tests, Primary I Battery, Form A (Bixler, Durost, Hildreth, Lund, & Wrightstone, 1958–1962). This group-administered battery includes four tests: Word Knowledge, Word Discrimination, Reading Comprehension, and Arithmetic.

Since NFT children took school-administered Metropolitan Achievement Tests twice a year during the last two years of the study, they became considerably more familiar with these instruments than did the FT children. For this reason, a different instrument was employed for all children in Grade 3 to measure academic achievement—the Peabody Individual Achievement Tests (PIAT; Dunn & Markwardt, 1970). The PIAT is an individually administered battery of Mathematics, Reading Recognition, Reading Comprehension, Spelling, and General Information tests.

Intellectual abilities. Form B of the Peabody

Picture Vocabulary Test (PPVT; Dunn, 1965) was used to investigate verbal intellectual development throughout the four years of the study. The PPVT is an individually administered test which assesses verbal conceptual knowledge independent of reading ability. In addition, the individually administered Picture Arrangement Test of the Wechsler Intelligence Scale for Children (Wechsler, 1949) was administered in Grade 3 to assess the children's nonverbal, problem-solving abilities.

Problem-solving style. Several measures were administered in first grade to assess the children's effectiveness in handling unfamiliar tasks. One of these measures, the Sticker Game (Zigler & Turnure, 1964), was designed to investigate the extent to which children rely on imitation rather than on their own ideas in carrying out a task. The Sticker Game yields imitation scores based on the similarity of designs the child makes to three designs constructed by the examiner. The frequency with which the children spontaneously engaged in verbal communication with the examiner was also recorded during the Sticker Game.

Also administered in first grade were the Circles Test from the Torrance Tests of Creative Thinking (Torrance, 1966), which was used to measure creativity on a nonverbal task, and a modified form of Torrance's Just Suppose Test, which was used to measure creativity on a verbal task. These tests assess creativity by the originality, fluency, and flexibility of ideas expressed by the child. The Sticker Game and Circles Test were individually administered to each child in one session. The Just Suppose Test was individually administered in a second session.

The Picture Arrangement Test employed in Grade 3 to measure nonverbal, problem-solving ability also provided an index of the degree to which children were impulsive as opposed to reflective when tackling new problems. Impulsivity was gauged by the amount of time children spent studying the pictures before attempting to arrange them in the correct order. Since the Picture Arrangement Test problems differ in difficulty and in time allowed to complete them, a relative time measure was used, a latency score consisting of the ratio of the preliminary study time (time from the "start" signal to the first reordering of the pictures) to the overall problem-solving time. A latency score of 15% thus indicated that a child spent 15% of his problem-solving time in preliminary study of the pictures. Lower latency scores imply a more impulsive approach; higher latency scores imply a more reflective approach.

Attitude toward school. Children's attitudes toward school were assessed in Grade 3 with the Attitude Toward School Questionnaire (ASQ; Klein & Strickland, 1970). The children were presented 54 situations involving classmates, teachers, lessons, and other aspects of school life. By circling one of three faces in their test booklets, the children indicated whether they would feel happy, neutral or don't know, or unhappy in each situation. The ASQ was group-administered in two sessions to each class. Responses to the 53 scored items (the first item was a practice item) were quantified on a scale of 1 equals unfavorable attitude, 2 equals neutral, and 3 equals favorable attitude. Attitude scores could thus range from 53 (totally unfavorable) to 159 (totally favorable).

The attitude data were factor analyzed to determine whether there were meaningful item clusters. Two main factors were revealed: an eight-item factor in which the items all concerned schoolwork and a seven-item factor in which all items involved situations where children were in contact with a teacher or principal. The subscores on these two factors (attitude toward academic work and attitude toward authorities) and the overall attitude scores were analyzed.

Self-image. A 32-item form of Coppersmith's Self-Esteem Inventory (Coopersmith, 1967) was administered in third grade, but the results are not reported because a considerable number of NFT children were unable to respond in a discriminating way to the Coopersmith items. After it became evident that Coopersmith results could not be used, a semiprojective, exploratory measure of children's perceptions of themselves as students in school was specially developed for this study. The instructions for this Student Self-Image measure (SSI), which was administered individually, were as follows:

> Now we're going to draw a picture and you have to pretend you are somewhere else. Here is a sheet of paper and a pencil. Pretend that you are in class. It is morning and the class is working. The picture could be called "[child's name] class at work." I'll write the title up here. Okay, you draw a picture here of your class at work... Now pretend the teacher said, "Today we are going to start on some new work. It is something we have never had before. Now listen and try to get it the first time." Would you probably get it the first time or not until later?... Pretend you're having trouble with the new work. Pretend you don't understand it. What would you do?... What would you do if you still didn't get it?... Do you think you'd ever get it?... Are there some kinds of work you do in school that you feel pretty good at?... What do you like best about school?

Three dimensions of the child's image of himself as a student were assessed on the SSI. The child's *sense of membership* in his class group was inferred from the manner in which he presented himself in his drawing of his class. Sense of membership was rated on a scale of 0–4 (0 = the child left himself out of the picture entirely; 1 = the child drew himself alone; 2 = the child drew himself with one other person; 3 = the child drew himself as part of a group; and 4 = the child drew himself as part of a group and the self-figure was given prominence by reason of size, detail, or central position).

The child's *feelings of effectiveness* in school were assessed from the confidence and initiative which he expressed in his responses to the ques-

tions concerning a new lesson. Confidence was rated as follows: 0, if the child did not think he would learn a new lesson the first time it was presented and thought he probably would never learn it; 1, if the child was uncertain about whether or not he would learn a new lesson; and 2, if the child was confident he would learn a new lesson and would probably learn it the first time it was presented. Initiative was rated as follows: 0, if the child could suggest no plan for coping with a learning difficulty in class; 1, if the child suggested asking someone else for help; and 2, if the child suggested a plan which entailed adaptive, self-initiated activity.

The child's *involvement in academic learning* was assessed from what he said he liked best about school. Learning involvement was rated as follows: 0, if the child said he liked nothing about school except lunch or recess; 1, if the child expressed a vague, general liking for school but not for any specific activity; 3, if the child said he most liked a nonacademic activity such as music or gym; and 4, if the child said he most liked an academic activity such as math, phonics, or reading. (No Number 2 rating was used in order to give this dimension the same range as the others.)

The ratings were summed to obtain an overall SSI score. These scores could range from 0 (negative self-image) to 12 (positive self-image). In order to insure impartial ratings, all identifying information was blocked out of the drawings and answer sheets. Each protocol was then scored independently by two raters, neither of whom had previously been associated with the study. The product-moment correlation coefficient for the two raters' SSI scores was +.87, indicating a satisfactorily high degree of interrater agreement. (Interrater agreement for the four individual ratings ranged from +.84 to +.95.)

Test behavior. A behavior checklist was employed to record each child's behavior during the two sessions in which the ASQ was administered in the third grade. Task-related verbal behavior of three types was recorded: (*a*) requests for assistance (e.g., "Can you erase this for me?"); (*b*) comments about procedures or materials (e.g., "How many pages are left?"); (*c*) comments or questions about the meaning of the test items (e.g., in response to the item, "You came to school in the morning. There is a sign near the door. It says 'No School Today.'" "What day is school closed? How I feel depends on what day school is closed."). Each type of verbal behavior was recorded as either present or absent during the presentation of each of the 54 ASQ items. Maladaptive behaviors (disruptions, withdrawal, refusal to do the task) were also recorded but are not reported because they occurred very infrequently.

Procedure

The testing schedule over the four years of the longitudinal study was as follows:

Kindergarten:
- September—PPVT (1 individual session)
- October—STAR (1 group session)
- April—STAR (1 group session)
- May—PPVT (1 individual session)

Grade 1:
- March—Problem-solving style (2 individual sessions)
- May—PPVT (1 individual session)
- June—Metropolitan Achievement Tests (2 group sessions)

Grade 3:
- March—ASQ, behavior checklist (2 group sessions)
- April—PPVT, SSI (1 individual session)
- May—Picture Arrangement Test, PIAT (1 individual session)

Testing of cross-sectional subjects was confined to Grade 3 and followed the Grade 3 schedule shown above.

Intelligence and academic achievement tests were administered in accordance with the standardized instructions given in the test manuals. The testing was carried out by 12 female examiners (4 for the kindergarten testing, 3 for the Grade 1 testing, and 5 for the Grade 3 testing). Group testing took place in the classrooms with 1 of the examiners presenting the test items and 2 or 3 others serving as proctors. (Teachers assisted the proctors, except during the Grade 3 administration of the ASQ when school staff were excluded from the classrooms.) Children were taken out of their classrooms for individual testing. The individual testing followed a schedule which insured that (*a*) every examiner tested both FT and NFT children on each measure and (*b*) no child was individually tested more than once by the same examiner.

Results

Analyses were performed on IQ, mental age (MA), and raw scores for the PPVT measure, but since in all cases the findings were the same for all three scores, only IQ scores are reported. For all other measures, analyses were performed on raw scores. Since the FT and NFT groups differed in their proportion of boys and girls, all analyses incorporated sex as a separate dimension.

Longitudinal Study

In order to assess the comparability of the FT and NFT groups at the outset of the longitudinal study, 2 × 2, Program (FT × NFT) × Sex unweighted means analyses of variance were conducted on the PPVT and STAR scores obtained at the beginning of kindergarten. No significant effects were

TABLE 3

PPVT Performance of Head Start Samples in FT and NFT Longitudinal Groups During Kindergarten and First Grade

Time of testing	FT - Head Start		NFT - Head Start		NFT - non-Head-Start	
	Boys	Girls	Boys	Girls	Boys	Girls
	(n = 20)	(n = 12)	(n = 4)	(n = 3)	(n = 8)	(n = 11)
Beginning of kindergarten	92.60	77.58	94.50	87.33	64.38	67.09
End of kindergarten	102.55	89.58	99.50	88.67	96.25	86.36
End of first grade	104.85	94.82	104.00	88.33	90.38	86.70

Note. Abbreviations: PPVT = Peabody Picture Vocabulary Test, FT = Follow Through, and NFT = non-Follow-Through.

found in the analysis of the PPVT scores. While no program effect was found for the STAR scores, a sex effect ($F = 8.47$, $df = 1/55$, $p < .01$) and a significant Program × Sex interaction ($F = 4.23$, $df = 1/55$, $p < .05$) were discovered. Comparisons among the four groups by the Newman-Keuls method (Winer, 1971) indicated that the sex effect was largely due to the high scores of the NFT girls.

As can be seen in Table 1, the FT and NFT groups differed in incidence of father absence and of Head Start experience.[5] The father absence variable was ignored after preliminary analyses revealed that it was not significantly related to any dependent measure at any testing.

Because of the extremely small number of non-Head-Start children in the FT program, the effects of Head Start experience were examined in 3 × 2, Education (FT – Head Start × NFT – Head Start × NFT – Non-Head-Start) × Sex unweighted means analyses of variance on each dependent measure at each testing.[6] Head Start effects were found only at the beginning of kindergarten and the end of first grade (see Table 3). Head Start experience had a significant effect on PPVT performance at the beginning of kindergarten (main effect for education, $F = 4.61$, $df = 2/52$, $p < .05$) when both FT and NFT Head Start groups scored higher than the NFT – non-Head-Start group. At the end of kindergarten, there were no differences in the PPVT performance of these three groups. The PPVT scores obtained at the end of first grade showed a combined Head Start – Follow Through effect. A main effect for education ($F = 3.47$, $df = 2/52$, $p < .05$) was found; however, comparisons among the groups revealed that only two of the three groups differed from each other: FT – Head Start children scored significantly higher than NFT – non-Head-Start children. Group comparisons indicated that FT – Head Start children also scored significantly higher than NFT – non-Head-Start children on the two verbal measures of problem-solving style administered in first grade (frequency of verbalization and creative thinking, verbal), although main effects for education were not found in the analyses of variance of these measures. No Head Start effects were found on any other kindergarten or Grade 1 measure nor were they found on any of the Grade 3 measures.

The scores of the FT and NFT children at every testing for school program effects are presented in Table 4. A 2 × 2 (Program × Sex) unweighted means analysis of variance was performed on each measure at each testing.[6] (One-tailed tests of statistical signifi-

[5] The groups also differed in proportion of blacks and whites, but there were too few white children in the sample to permit meaningful analyses for race effects.

[6] Covariance adjustment was possible for the PPVT measure since it was administered at every testing. End-of-kindergarten raw scores were used as the covariable for three-group comparisons (FT – Head Start × NFT – Head Start × NFT – Non-Head-Start) and two-group comparisons (FT × NFT) of Grade 1 raw scores and Grade 3 raw scores. (Beginning kindergarten scores could not be used because the PPVT performance of

TABLE 4

SCORES FOR FT AND NFT LONGITUDINAL GROUPS

Measure	Follow Through				Non-Follow-Through			
	Boys (n = 21)		Girls (n = 14)		Boys (n = 12)		Girls (n = 14)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
End of kindergarten								
PPVT	101.19	9.90	88.36	18.67	97.33	15.37	86.86	17.31
STAR	50.05	10.70	51.79	11.99	47.18	9.06	53.14	9.44
End of Grade 1								
PPVT	103.52	10.10	94.23	11.20	94.09	10.41	87.08	12.90
MAT, Word Knowledge	19.39	7.54	20.69	9.29	19.10	7.11	16.50	9.69
MAT, Word Discrimination	19.50	9.38	21.31	9.35	19.30	9.50	18.42	7.44
MAT, Reading Comprehension	15.06	6.95	20.17	10.11	13.67	5.77	14.92	8.12
MAT, Arithmetic	39.35	12.04	39.92	15.38	42.20	11.72	40.08	10.08
Imitation	9.70	5.25	9.38	4.56	10.73	5.44	12.50	5.23
Frequency of communication	7.85	9.44	9.54	14.16	2.82	4.19	1.64	2.34
Creative thinking, nonverbal	35.15	21.78	28.08	17.58	32.46	19.18	32.14	18.79
Creative thinking, verbal	9.10	5.97	7.62	6.28	6.18	6.51	4.36	4.77
End of Grade 3								
PPVT	100.95	11.51	89.57	10.46	92.92	9.10	86.07	12.55
Picture Arrangement Test	23.67	6.95	20.64	6.26	19.46	8.10	17.29	8.92
PIAT, Mathematics	37.05	8.29	33.57	8.51	32.91	10.42	29.93	7.71
PIAT, Reading Recognition	32.14	8.52	30.64	8.48	34.09	8.83	30.64	6.17
PIAT, Reading Comprehension	30.48	7.62	31.00	9.65	31.54	7.54	30.64	4.83
PIAT, Spelling	33.48	9.30	34.93	13.22	36.46	9.21	33.14	7.56
PIAT, General Information	26.81	8.69	19.00	8.37	18.82	7.51	19.36	6.61
Problem-solving latency	31.05	7.45	27.00	5.39	32.64	6.07	25.36	8.68
ASQ, academic work	15.94	4.28	17.15	4.24	16.58	5.00	18.00	3.02
ASQ, school authorities	12.59	3.16	13.08	2.96	12.75	2.56	13.08	2.58
ASQ, overall attitude	114.71	8.64	119.08	13.09	115.25	16.09	123.17	6.47
Student Self-Image	8.00	3.12	8.00	2.63	6.50	3.00	8.75	2.80
Comments on procedures	11.33	10.02	7.62	7.18	7.25	6.94	3.57	4.38
Requests for help	.86	1.42	1.92	2.02	1.00	1.48	.64	1.65
Comments on meaning	2.10	3.08	1.92	2.63	.50	.91	.29	.61

Note. Abbreviations: FT = Follow Through, NFT = non-Follow-Through, PPVT = Peabody Picture Vocabulary Test, STAR = Screening Test of Academic Readiness, MAT = Metropolitan Achievement Test, PIAT = Peabody Individual Achievement Test, and ASQ = Attitude Toward School Questionnaire.

cance were employed to assess program main effects in view of the expectation that the FT children would have higher scores. Two-tailed tests were employed to assess sex main effects and interaction effects.) The results of these analyses are summarized in Table 5. The three program effects found at the end of Grade 1 were combination Head Start – Follow Through effects, as reported above. On every Grade 3 measure where a program effect was found, FT children scored higher than both NFT – Head Start and NFT – non-Head-Start children.

Cross-Sectional Study

Preliminary analyses of the Grade 3 cross-sectional data revealed that father absence and Head Start attendance did not significantly influence performance on any of the dependent measures. The scores of the FT and NFT children are presented in Table 6. A 2 × 2 (Program × Sex) unweighted means analysis of variance was conducted on each of the measures. The results of these analyses are summarized in Table 7.

many children appeared to be attenuated by motivational factors at that testing.) The findings were comparable to those reported for the unadjusted analyses of variance except that the significance levels for the Grade 1 differences were improved.

TABLE 5
SUMMARY OF FINDINGS FOR FT AND NFT LONGITUDINAL GROUPS

Measure	Program	Sex	Program × Sex
End of kindergarten			
PPVT	*ns*	boys > girls*	*ns*
STAR	*ns*	*ns*	*ns*
End of Grade 1			
PPVT	FT > NFT**	boys > girls**	*ns*
MAT, Word Knowledge	*ns*	*ns*	*ns*
MAT, Word Discrimination	*ns*	*ns*	*ns*
MAT, Reading Comprehension	*ns*	*ns*	*ns*
MAT, Arithmetic	*ns*	*ns*	*ns*
Imitation	*ns*	*ns*	*ns*
Frequency of communication	FT > NFT**	*ns*	*ns*
Creative thinking, nonverbal	*ns*	*ns*	*ns*
Creative thinking, verbal	FT > NFT*	*ns*	*ns*
End of Grade 3			
PPVT	FT > NFT*	boys > girls**	*ns*
Picture Arrangement Test	FT > NFT*	*ns*	*ns*
PIAT, Mathematics	FT > NFT*	*ns*	*ns*
PIAT, Reading Recognition	*ns*	*ns*	*ns*
PIAT, Reading Comprehension	*ns*	*ns*	*ns*
PIAT, Spelling	*ns*	*ns*	*ns*
PIAT, General Information	FT > NFT*	boys > girls*	*ns*
Problem-solving latency	*ns*	boys > girls**	*ns*
ASQ, academic work	*ns*	*ns*	*ns*
ASQ, school authorities	*ns*	*ns*	*ns*
ASQ, overall attitude	*ns*	*ns*	*ns*
Student Self-Image	*ns*	*ns*	*ns*
Comments on procedures	FT > NFT*	*ns*	*ns*
Requests for help	*ns*	*ns*	*ns*
Comments on meaning	FT > NFT**	*ns*	*ns*

Note. Abbreviations: FT = Follow Through, NFT = non-Follow-Through, PPVT = Peabody Picture Vocabulary Test, STAR = Screening Test of Academic Readiness, MAT = Metropolitan Achievement Test, PIAT = Peabody Individual Achievement Test, and ASQ = Attitude Toward School Questionnaire.

* $p < .05$.

** $p < .01$.

Since the NFT children attended three types of school program (mixed socioeconomic grouping with enriched program, tutorial program, traditional program), a 3 × 2, School (Mixed × Tutorial × Traditional) × Sex unweighted means analysis of variance was performed on the NFT children's scores on each of the dependent measures. The results are summarized in Table 8. On 2 of the 11 measures on which a significant FT effect was found (Mathematics and problem-solving latency), a significant difference among the three NFT schools was also found. On these 2 measures, the analysis was rerun with the FT children included as a fourth school group. On the Mathematics test, a school main effect ($F = 6.03$, $df = 3/132$, $p < .001$) was again found. Employing the Newman-Keuls procedure, comparisons of the four schools indicated that FT children scored significantly higher than NFT children in the tutorial and traditional schools, but FT children did not score differently from NFT children in the socioeconomically mixed school. On the problem-solving latency measure, a school main effect ($F = 7.50$, $df = 3/132$, $p < .001$) was again found. Comparisons of the four schools indicated that FT children had significantly higher latency scores than NFT children in only one of the NFT schools, the traditional school.

In the major cross-sectional analyses, then, FT children were compared with a

TABLE 6

SCORES FOR FT AND NFT CROSS-SECTIONAL GROUPS AT END OF THIRD GRADE

Measure	Follow Through				Non-Follow-Through			
	Boys (n = 25)		Girls (n = 17)		Boys (n = 51)		Girls (n = 49)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
PPVT	101.84	11.10	90.24	9.69	91.26	12.33	86.41	11.04
Picture Arrangement Test	23.92	6.63	20.88	6.18	19.41	7.72	16.61	7.65
PIAT, Mathematics	36.48	8.36	35.29	8.62	33.27	9.56	28.29	7.98
PIAT, Reading Recognition	32.64	9.00	31.65	8.25	30.78	8.41	30.61	6.36
PIAT, Reading Comprehension	30.80	7.91	31.88	9.17	29.18	7.05	29.14	5.41
PIAT, Spelling	33.80	9.18	36.18	12.78	32.12	8.36	32.22	6.95
PIAT, General Information	27.84	9.94	20.88	9.95	20.16	7.76	17.74	5.77
Problem-solving latency	31.20	7.82	27.94	5.36	26.76	9.68	23.35	8.99
ASQ, academic work	15.96	4.10	17.63	3.95	17.11	4.21	18.29	3.41
ASQ, school authorities	12.92	3.59	13.38	3.05	12.69	3.62	12.67	3.66
ASQ, overall attitude	114.56	10.36	120.25	12.44	116.54	15.19	122.23	10.91
Student Self-Image	8.12	2.85	8.06	2.44	6.77	2.35	7.41	2.95
Comments on procedures	12.36	9.81	7.25	7.25	7.17	6.75	3.92	4.18
Requests for help	.80	1.35	2.50	3.37	1.15	1.73	.75	1.50
Comments on meaning	3.16	4.49	2.38	2.90	.22	.59	.46	1.34

Note. Abbreviations: FT = Follow Through, NFT = non-Follow-Through, PPVT = Peabody Picture Vocabulary Test, PIAT = Peabody Individual Achievement Test, and ASQ = Attitude Toward School Questionnaire.

TABLE 7

SUMMARY OF FINDINGS FOR FT AND NFT CROSS-SECTIONAL GROUPS AT END OF THIRD GRADE

Measure	Program	Sex	Program × Sex
PPVT	FT > NFT**	boys > girls***	*ns*
Picture Arrangement Test	FT > NFT**	boys > girls*	*ns*
PIAT, Mathematics	FT > NFT**	*ns*	*ns*
PIAT, Reading Recognition	*ns*	*ns*	*ns*
PIAT, Reading Comprehension	FT > NFT*	*ns*	*ns*
PIAT, Spelling	FT > NFT*	*ns*	*ns*
PIAT, General Information	FT > NFT***	boys > girls**	*ns*
Problem-solving latency	FT > NFT**	boys > girls*	*ns*
ASQ, academic work	*ns*	*ns*	*ns*
ASQ, school authorities	*ns*	*ns*	*ns*
ASQ, overall attitude	*ns*	girls > boys*	*ns*
Student Self-Image	FT > NFT*	*ns*	*ns*
Comments on procedures	FT > NFT**	boys > girls**	*ns*
Requests for help	FT > NFT*	*ns*	FT girls > NFT boys > FT boys NFT girls**
Comments on meaning	FT > NFT***	*ns*	*ns*

Note. Abbreviations: FT = Follow Through, NFT = non-Follow-Through, PPVT = Peabody Picture Vocabulary Test, PIAT = Peabody Individual Achievement Test, and ASQ = Attitude Toward School Questionnaire.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

TABLE 8
SUMMARY OF FINDINGS FOR THREE NFT SCHOOLS (MIXED SOCIOECONOMIC STATUS, TUTORIAL, TRADITIONAL) IN CROSS-SECTIONAL GROUP AT END OF THIRD GRADE

Measure	School	Sex	School × Sex
PPVT	*ns*	*ns*	*ns*
Picture Arrangement Test	*ns*	*ns*	T boys & M girls > others*
PIAT, Mathematics	M & T > Tr*	boys > girls**	*ns*
PIAT, Reading Recognition	*ns*	*ns*	*ns*
PIAT, Reading Comprehension	*ns*	*ns*	*ns*
PIAT, Spelling	*ns*	*ns*	*ns*
PIAT, General Information	*ns*	*ns*	*ns*
Problem-solving latency	M & T > Tr**	*ns*	*ns*
ASQ, academic work	*ns*	*ns*	*ns*
ASQ, school authorities	M > T > Tr**	*ns*	M girls > others > Tr girls**
ASQ, overall attitude	*ns*	girls > boys*	*ns*
Student Self-Image	*ns*	*ns*	*ns*
Comments on procedures	*ns*	boys > girls**	*ns*
Requests for help	*ns*	*ns*	*ns*
Comments on meaning	*ns*	*ns*	*ns*

Note. Abbreviations: NFT = non-Follow-Through, M = mixed socioeconomic status, T = tutorial, Tr = traditional, PPVT = Peabody Picture Vocabulary Test, PIAT = Peabody Individual Achievement Test, and ASQ = Attitude Toward School Questionnaire.
* $p < .05$.
** $p < .01$.

group of NFT children whose performance on these two measures was influenced by the program of the particular school they were attending.

Comparisons of FT Economically Disadvantaged and FT Nondisadvantaged Children

The scores of the FT nondisadvantaged children at every testing are presented in Table 9. At the outset of the program (beginning kindergarten), these children scored higher than the FT economically disadvantaged children on both the PPVT ($F = 8.84$, $df = 1/44$, $p < .01$) and STAR ($F = 9.99$, $df = 1/42$, $p < .01$). Analyses revealed no significant changes in the nondisadvantaged children's PPVT IQs over the course of the four-year program. A 2 × 2, Income Group (Disadvantaged × Nondisadvantaged) × Sex unweighted means analysis of variance was conducted on every measure at every testing which assessed the effects of the FT program. The results of these analyses are summarized in Table 10.

DISCUSSION

Effects of Head Start

Given the very small number of FT children who had not had Head Start, essentially what was examined in this study was the impact of a five-year compensatory effort—one year of Head Start and four years of Follow Through. Consistent with current views, the findings of the present study indicate that a one-year Head Start experience, if not followed by a subsequent compensatory effort, has little lasting effect on children's performance. The Head Start effects that were found were limited to the beginning of kindergarten and end of first grade, and these pose certain problems of interpretation. At the beginning of kindergarten, the PPVT effect was due to the inordinately low scores obtained by non-Head-Start children. This effect disappeared at the end of kindergarten when the non-Head-Start children showed a 25-point IQ increase. In view of the evidence (Beller, 1969; Thomas, Hertzig, Dryman, & Fernandez, 1971; Zigler, Abelson, & Seitz, 1973; Zigler & Butterfield, 1968) that deleterious motivational factors attenuate the intelligence test performance of disadvantaged children, it seems probable from the large IQ increase of the non-Head-Start children that their low PPVT scores at school entrance were due to such factors rather than being accurate indicators of their actual cognitive abilities.

TABLE 9

SCORES FOR FT NONDISADVANTAGED BOYS AND GIRLS

Measure	Boys (n = 5)		Girls (n = 5)	
	M	*SD*	*M*	*SD*
End of kindergarten				
PPVT	120.20	9.73	101.60	21.69
STAR	62.00	8.83	65.50	3.70
End of Grade 1				
PPVT	126.00	28.28	105.50	15.16
MAT, Word Knowledge	29.00	6.24	27.00	7.52
MAT, Word Discrimination	30.33	5.03	27.60	7.89
MAT, Reading Comprehension	31.00	18.52	29.50	14.06
MAT, Arithmetic	46.33	20.11	42.60	10.67
Imitation	9.75	5.56	10.20	6.22
Frequency of communication	22.50	24.09	8.60	7.80
Creative thinking, nonverbal	52.75	15.31	20.00	13.78
Creative thinking, verbal	—[a]	—[a]	7.00	3.83
End of Grade 3				
PPVT	116.60	9.92	97.80	11.01
Picture Arrangement Test	31.00	7.87	23.00	5.83
PIAT, Mathematics	49.60	14.98	34.60	11.33
PIAT, Reading Recognition	40.80	17.24	36.20	7.36
PIAT, Reading Comprehension	37.00	12.90	32.80	7.60
PIAT, Spelling	41.40	14.66	42.80	13.54
PIAT, General Information	43.60	13.65	28.60	10.94
Problem-solving latency	25.40	5.46	25.80	8.29
ASQ, academic work	16.40	4.04	17.25	3.40
ASQ, school authorities	12.80	1.64	9.50	1.92
ASQ, overall attitude	111.40	8.26	115.00	16.31
Student Self-Image	8.20	2.59	8.80	2.95
Comments on procedures	9.60	6.15	1.75	2.36
Requests for help	.20	.45	.00	.00
Comments on meaning	4.40	3.29	.00	.00

Note. Abbreviations: FT = Follow Through, PPVT = Peabody Picture Vocabulary Test, STAR = Screening Test of Academic Readiness, MAT = Metropolitan Achievement Test, PIAT = Peabody Individual Achievement Test, and ASQ = Attitude Toward School Questionnaire.

[a] Scores were obtained for only two nondisadvantaged boys on this measure.

The reemergence of a significant PPVT performance difference at the end of first grade between FT – Head Start children and NFT – non-Head-Start children was partly due to the gains made by FT children and partly due to losses by NFT – non-Head-Start boys. These boys seem not to have profited as much from their schooling—at least as measured by PPVT performance—as the other children in NFT schools. The poorer performance of these boys again may have been due to motivational factors, in that it was evident that NFT first-grade classrooms did not provide as warm and supportive an environment for the children as did NFT kindergartens.

The findings for the first-grade, verbal problem-solving style measures show the same transitional pattern as the PPVT IQ results. The FT children were generally ahead of NFT children in verbal performance abilities, but their superiority was clear-cut only in comparison to NFT children who had not had Head Start. Two years later this was no longer the case. On all Grade 3 measures where FT children were superior, they were superior to both the Head Start and non-Head-Start children in NFT schools.

Effects of Follow Through

The findings at the end of third grade showed a consistent pattern of superior performance by FT over NFT children. In the longitudinal study, the FT children were superior on such intelligence–achievement

TABLE 10

SUMMARY OF FINDINGS FOR ECONOMICALLY DISADVANTAGED (D) AND NONDISADVANTAGED (ND) CHILDREN IN FOLLOW THROUGH

Measure	Income group	Sex	Income Group × Sex
End of kindergarten			
PPVT	ND > D**	boys > girls**	*ns*
STAR	ND > D**	*ns*	*ns*
End of Grade 1			
PPVT	ND > D***	boys > girls***	ND boys > D boys / ND girls > D girls*
MAT, Word Knowledge	ND > D*	*ns*	*ns*
MAT, Word Discrimination	ND > D*	*ns*	*ns*
MAT, Reading Comprehension	ND > D**	*ns*	*ns*
MAT, Arithmetic	*ns*	*ns*	*ns*
Imitation	*ns*	*ns*	*ns*
Frequency of communication	*ns*	*ns*	*ns*
Creative thinking, nonverbal	*ns*	boys > girls**	*ns*
Creative thinking, verbal	—[a]	—[a]	—[a]
End of Grade 3			
PPVT	ND > D*	boys > girls***	*ns*
Picture Arrangement Test	*ns*	boys > girls*	*ns*
PIAT, Mathematics	*ns*	boys > girls*	*ns*
PIAT, Reading Recognition	ND > D*	*ns*	*ns*
PIAT, Reading Comprehension	*ns*	*ns*	*ns*
PIAT, Spelling	*ns*	*ns*	*ns*
PIAT, General Information	ND > D***	boys > girls**	*ns*
Problem-solving latency	*ns*	*ns*	*ns*
ASQ, academic work	*ns*	*ns*	*ns*
ASQ, school authorities	*ns*	*ns*	*ns*
ASQ, overall attitude	*ns*	*ns*	*ns*
Student Self-Image	*ns*	*ns*	*ns*
Comments on procedures	*ns*	*ns*	*ns*
Requests for help	D > ND*	*ns*	*ns*
Comments on meaning	*ns*	boys > girls*	*ns*

Note. Abbreviations: PPVT = Peabody Picture Vocabulary Test, STAR = Screening Test of Academic Readiness, MAT = Metropolitan Achievement Test, PIAT = Peabody Individual Achievement Test, and ASQ = Attitude Toward School Questionnaire.

[a] This analysis was not performed because scores were available for only two nondisadvantaged boys.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

measures as the PPVT, the Picture Arrangement Test, Mathematics, and General Information. In regard to social-motivational characteristics, the FT children demonstrated greater curiosity than the NFT children by commenting more frequently about the meaning and significance of tasks presented to them. These same differences favoring FT children were discovered in the cross-sectional study. In addition, in the cross-sectional study, the FT children were superior to NFT children on the intelligence–achievement measures of Reading Comprehension and Spelling. The FT children also displayed a more reflective approach to problem solving, a greater willing-

ness to ask adults for assistance, and a more positive image of themselves as students.[7]

It was particularly noteworthy to find that a Bank Street approach, which emphasizes the discovery method, was associated with superior performance not only on intelligence and social-motivational measures but on academic achievement measures as well. National assessment findings (Stanford Research Institute, 1971) suggested that discovery approaches were not as effective as more structured curriculum approaches in producing gains on achievement measures. These national findings, which were based on assessments after a year of school attendance, are consistent only with the kindergarten and Grade 1 findings of the present study indicating no differences between FT and NFT children on the kindergarten Screening Test of Academic Readiness and the Grade 1 Metropolitan Achievement Tests. However, FT children were superior to NFT children on achievement measures after FT children had experienced a Bank-Street-type program for four years. There is no way of determining from the present study whether even better achievement performance would accrue with a more structured approach. It is important to note, though, that these third-grade findings do not support the view which has developed (Pines, 1967) that the "whole child" approach does not effectively improve the achievement scores of disadvantaged children.

The findings of this study must be interpreted with some caution in view of the fact that children were not randomly assigned to FT and NFT programs at the beginning of the study. Whether there were selective factors operating and in what direction such factors might have influenced the results cannot be assessed until replication studies are completed in which there has been random assignment of children.[8] (Such a study is now in progress at this FT program.) It was nevertheless encouraging that positive effects were visible in the FT children studied here. The findings certainly call into question the general proposition that compensatory education is a failure. In addition to demonstrating that children who received a special program for several years were superior on a wide array of measures to children who did not receive such a program, the findings also suggest that other types of remedial efforts benefit disadvantaged children. Comparison children attending a school with a mixed socioeconomic grouping and enriched program and a school with a tutorial program displayed a less impulsive approach to problem solving, greater mathematics achievement, and a more positive attitude toward school authorities than did children who were attending a school which had no special remedial or enrichment programs of any kind.

In interpreting the Follow Through program effects, it is important that the FT children not only had a prior year of Head Start experience but that during their four years in the program they were in interaction with a number of children from better-educated, higher-income families. The finding of better performance in the FT and NFT schools with mixed socioeconomic and racial groupings is consistent with the findings of the Coleman report (Coleman, 1966). This finding also supports the view recently expressed by defenders of busing (Pettigrew, Useem, Normand, & Smith, 1973) that in order for achievement to be higher in desegregated schools, other crucial conditions (such as adequate school services and remedial training) must be met, as they were in this FT program. Just how important socioeconomic mixing was to this program will be clarified when studies are completed of similar FT programs attended only by economically disadvantaged children.

Economically Disadvantaged and Nondisadvantaged Comparisons

Although there was no nondisadvantaged non-FT group with which to compare the FT nondisadvantaged children at the end of third grade, there is no evidence that attend-

[7] The similarity in the test scores of the NFT cross-sectional and NFT longitudinal groups suggests that the performance of the cross-sectional children was not impaired by having less testing experience.

[8] A penetrating discussion of these issues may be found in Campbell and Erlebacher (1970).

ing a program in which disadvantaged children were in the majority was detrimental to the learning growth of the nondisadvantaged children. These children's PPVT IQ scores did not change over their four years in the program. Furthermore, converting the nondisadvantaged children's scores on the third-grade achievement tests into grade equivalent levels indicated that these children were performing at or (more typically) above grade level in all academic areas.

The nondisadvantaged children attending the Follow Through program were generally superior to the economically disadvantaged children in the program on intelligence and achievement measures at all testings. This superiority was less marked in the fourth year of school than it was in the first two years. At the end of Grade 3, no significant differences were found between the economically disadvantaged and nondisadvantaged children on the Picture Arrangement, Mathematics, Reading Comprehension, or Spelling tests, or on any of the social-motivational measures. It would be erroneous to conclude from this that the FT program resulted in the disadvantaged children attaining a level of intellectual functioning equivalent to that of the nondisadvantaged children, since on every measure where a difference was found, the difference was in favor of the nondisadvantaged children. But, given the striking differences in the home lives of these children, it would surely be naive to expect that a compensatory program alone would be powerful enough to produce equivalent intellectual functioning.

In sum, the findings of this study indicate that while the Follow Through program assessed was not capable of ameliorating all of the negative effects of living in an economically disadvantaged environment, the program was highly beneficial to the children who participated in it. The longitudinal and cross-sectional evidence together point to the conclusion that the gains accruing from compensatory education programs are commensurate with the duration and amount of effort which are expended on these programs.

REFERENCES

Abelson, W. D. Head Start graduates in school: Studies in New Haven, Connecticut. In S. Ryan (Ed.), *A report on longitudinal evaluations of preschool programs*. Vol. 1. *Longitudinal evaluations*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1974.

Ahr, A. E. *Screening test of academic readiness*. Skokie, Ill.: Priority Innovations, 1966.

Beller, E. K. The evaluation of effects of early educational intervention on intellectual and social development of lower-class, disadvantaged children. In, *Critical issues in research related to disadvantaged children*. Princeton, N. J.: Educational Testing Service, 1969.

Bixler, H. H., Durost, W. N., Hildreth, G. T., Lund, K. W., & Wrightstone, J. W. *Metropolitan Achievement Test series*. New York: Harcourt, Brace, & World, 1958–1962.

Bronfenbrenner, U. *A report on longitudinal evaluations of preschool programs*. Vol. 2. *Is early intervention effective?* Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1974.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Disadvantaged child*. Vol. 3. New York: Brunner/Mazel, 1970.

Coleman, J. S. *Equality of educational opportunity*. Washington, D. C.: U. S. Government Printing Office, 1966.

Coopersmith, S. *The antecedents of self-esteem*. San Francisco: W. H. Freeman, 1967.

Dunn, L. M. *Peabody Picture Vocabulary Test*. Minneapolis, Minn.: American Guidance Service, 1965.

Dunn, L. M. & Markwardt, F. C. *Peabody Individual Achievement Tests*. Circle Pines, Minn.: American Guidance Service, 1970.

Eysenck, H. J. *The IQ argument*. New York: Library Press, 1971.

Jensen, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 1969, **39**, 1–123.

Klaus, R. A., & Gray, S. W. The early training project for disadvantaged children: A report after five years. *Monographs of the Society for Research in Child Development*, 1968, **33**(4), 1–66.

Klein, S. P., & Strickland, G. *The child's attitude toward school*. Los Angeles, University of California, Center for the Study of Evaluation, 1970. (ERIC Document Reproduction Service No. Ed 048 324)

Kohlberg, L. Early education: A cognitive-development view. *Child Development*, 1968, **39**, 1013–1062.

Maccoby, E., & Zellner, M. *Experiments in primary education: Aspects of Project Follow-Through*. New York: Harcourt-Brace, 1970.

Pettigrew, T. F., Useem, E. L., Normand, C., &

Smith, M. S. Busing: A review of "the evidence." *Public Interest*, 1973, **30**, 88–118.

Pines, M. *Revolution in learning*. New York: Harper & Row, 1967.

Ryan, S. (Ed.) *A report on longitudinal evaluations of preschool programs*. Vol. 1. *Longitudinal evaluations*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1974.

Stanford Research Institute. *Longitudinal evaluation of selected features of the national Follow Through program*. Menlo Park, Calif.: Author, 1971, (ERIC Document Reproduction Service Nos. Ed 057 266 and Ed 057 267)

Thomas, A., Hertzig, M. E., Dryman, I., & Fernandez, P. Examiner effect in IQ testing of Puerto Rican working-class children. *American Journal of Orthopsychiatry*, 1971, **41**, 809–821.

Torrance, E. P. *Torrance Tests of Creative Thinking*. Princeton, N. J.: Personnel Press, 1966.

Wechsler, D. *Wechsler Intelligence Scale for Children*. New York: Psychological Corporation, 1949.

Weikart, D. B. Early childhood special education for intellectually subnormal and/or culturally different children. Manuscript prepared for the National Leadership Institute in Early Childhood Development in Washington, D. C. Ypsilanti, Mich.: High/Scope Educational Research Foundation, 1971.

Westinghouse Learning Corporation. *The impact of Head Start experience: An evaluation of the effects of Head Start on Children's cognitive and affective development*. Vol. 1. *Text and Appendices A–E*. Ohio University Report to Office of Educational Opportunity, Clearinghouse for Federal, Scientific, and Technical Information, Washington, D. C., 1969.

Winer, B. J. *Statistical principles in experimental design*. (2nd. ed.) New York: McGraw-Hill, 1971.

Zigler, E. Project Head Start: Success or failure? *Learning*, 1973, **1**, 43–47.

Zigler, E., Abelson, W. D., & Seitz, V. Motivational factors in the performance of economically disadvantaged children on the Peabody Picture Vocabulary Test. *Child Development*, 1973, **44**, 294–303.

Zigler, E., & Butterfield, E. C. Motivational aspects of changes in IQ test performance of culturally deprived nursery school children. *Child Development*, 1968, **39**, 1–14.

Zigler, E., & Turnure, J. Outer-directedness in the problem solving of normal and retarded children. *Journal of Abnormal and Social Psychology*, 1964, **69**, 427–436.

(Received November 27, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 772-779

ABSTRACTIVE PROCESSES IN THE REMEMBERING OF PROSE[1]

RONALD E. JOHNSON[2]

Purdue University

The patterning of recall of linguistic subunits of textual prose was found to be strongly related to the semantic dimensions of abstractness–concreteness, specificity of denotation, comprehensibility, and interest. Attesting to the generality of the findings, the relationships between the textual dimensions and recall were evident with two textual passages, with two methods of measuring the semantic dimensions, and at immediate and delayed retention intervals. Remembering was characterized as being both a reconstructive and an abstractive process.

The remembering of semantic information from prose appears not to be based upon a literal remembering of exact words (Bartlett, 1932). Verbatim recalls are the exception, and the usual mode of recall is the paraphrase. As the retention interval lengthens, correct paraphrasings also decline, and the number of qualitative changes in remembering increases (Johnson, 1962). Since semantic units are amalgamated in recall, Cofer (1973) reaffirms Bartlett's (1932) theory that memories are stored in schematic form and that recalls are constructions from the schemas.

Zangwill (1972), in contrast, asserts that "remembering is better described as an abstractive than as a constructive process [p. 128]." Although Zangwill acknowledged qualitative changes, the predominant change in recall was noted to be omissions (Gomulicki, 1956; Johnson, 1962). More important, the patterning of omissions was not random, in that the prose elements which carried action content were remembered quite well whereas other segments were remembered poorly. Since the learner had to process all subunits in order to emphasize selected subunits, Zangwill (1972) argued that the occurrence of *selective* omissions reflects an abstractive process in learning.

What are the subunits in prose which are likely to be remembered? For narrative prose, the remembering of individual subunits is positively related to the importance of the subunits to the semantic structure of the passage (Gomulicki, 1956; Johnson, 1970). For textual prose, segments judged important in the logical structure of a passage were remembered better (Meyer & McConkie, 1973). Variations in meaningfulness also are strongly related to the remembering of textual subunits (Johnson, 1973).

The present research explored the possibility that additional textual dimensions were related to the patterning of omissions. For the dimensions of interest, concreteness, specificity, and comprehensibility, previous studies of prose learning have not examined the possibility that dimensional variations among prose subunits may be related to the patternings of recall. The present study attempts to fill this empirical gap.

METHOD

Materials

Two textual passages were adapted from college textbooks. One passage, entitled "The Role of Language in Learning," was a modified excerpt from Carroll's (1964) *Language and Thought*.[3] "Language" contained 650 words in 5 paragraphs

[1] The specificity data were presented in a paper delivered at the 1972 meeting of the Midwestern Psychological Association, Cleveland. Analyses of the concreteness data were presented at the 1972 meeting of the American Psychological Association, Honolulu.

[2] Requests for reprints should be sent to Ronald E. Johnson, SCC "G," Purdue University, West Lafayette, Indiana 47907.

[3] Copies of the textual passages and the instructions given to the various raters may be obtained from the author.

and 19 sentences. The second passage, "Evolution of the Brain," was derived from prose excerpts in introductory psychology textbooks written by Munn (1946) and Wickens and Meyer (1955). "Evolution" had 810 words in its 7 paragraphs and 40 sentences.

A sample of 52 college students, enrolled in psychology courses, divided "Language" into prose subunits by indicating locations which were psychologically acceptable for pausing (Johnson, 1970). The validity of a psycholinguistic boundary was assumed when a majority of raters agreed that the junction was an acceptable location for pausing. Since some subunits were only one or two words in length, the experimenter reduced the resulting 69 units to 60 units by combining introductory sentence modifiers with their clauses and by combining units in seriation. The 60 subunits had a mean word length of 10.8 (SD = 4.8). Seven of the subunits were intact sentences. Twenty of the pausal boundaries were located at the junctions between sentences, while 41 boundaries occurred within sentences. In a determination of the reliability of the segmentation process, the 52 raters were divided randomly into two groups of 26 raters, and the pausal locations in the two sets of ratings showed a Pearson correlation of .98.

Through the same partitioning process used with "Language," 50 of the 52 raters segmented "Evolution" into 89 pausal units. After the experimenter combined short introductory modifiers and serial units into adjacent subunits, the number of prose subunits was 80. Forty pausal boundaries were located within sentences, and 11 of the sentences were unbroken by a pausal boundary. The mean number of words in each subunit was 10.1 (SD = 5.3). A random split of the 50 raters into two groups of 25 raters again showed high reliability for the rater judgments (r = .97).

Dimensional Ratings

Raters of abstractness–concreteness were told that concrete phrases, such as "the wounded fox was relentlessly pursued by the barking hounds," had ready reference to concrete, tangible objects or events. In contrast, an abstract phrase, such as "the contemporary influence had not been anticipated previously," was described as more general and less readily referenced to particular persons, places, or things.

Specificity was to be judged by several criteria. First, the generality of the words themselves, for example, *furniture* versus *table*, was a determinant of the generality of the prose segment. Second, phrases telling what something was, for example, *wrinkled*, were to be judged as more specific than phrases telling what something was not, for example, *not very smooth*. Third, the presence of more detail in a description ordinarily was associated with greater specificity. Raters were advised, however, that the length of the phrase was not a reliable indicator of the amount of detail. Fourth, phrases referring to an exact quantity were said to be more specific, for example, *six bottles* as opposed to *many bottles*. Fifth, nouns modified by adjectives, and verbs modified by adverbs were to be judged as more specific than their unmodified counterparts. Sixth, phrases that included prepositional phrases were said to be more likely to be high in specificity, for example, *car with a dent in the fender* versus *car with a dent* versus *car*.

Differences in comprehensibility were said to be evident by the ease with which phrases could be apprehended or understood. Variations in comprehensibility were said to be most evident during the initial reading of passages and also during rapid readings. Acknowledging that most phrases in textual passages could be apprehended without too much difficulty, it was suggested that raters judge how comprehensible the phrases would be to students who were either younger or less intelligent than themselves.

Raters of interest were told that within any textual passage, including dull passages, some portions of content were more interesting than others. Judgments regarding interest, therefore, were to be made in the context of how interesting a phrase was in comparison with other phrases in the same textual passage.

Each textual dimension was assessed by two methods of rating. In the eliminative method (Johnson, 1970), raters judged the prose units by eliminating units which were lowest on the dimension being rated. A count of the number of times that a subunit was allowed to remain in the passage provided a measure of the unit's relative status on that particular dimension. Subunits were eliminated qua subunits, and the total word count for all remaining subunits had to be within 15 words of a specified number. Equal numbers of raters were assigned to groups eliminating either one fourth, one half, or three fourths of the words. Each rater judged both textual passages and eliminated the same proportion of words in each passage. The raters' passages had diagonal slashes separating the subunits, and word counts were printed in the margins.

Raters were recruited from introductory classes in psychology, and they performed the ratings in exchange for experimental credit. Independent groups of subjects rated each of the four dimensions (numbers of raters making eliminative judgments: concreteness, n = 72; specificity, n = 75; comprehensibility, n = 78; and interest, n = 99).

The textual dimensions also were assessed by 7-point categorical rating scales. Prior to performing the ratings, the raters read the passage in its usual prose format. The rating form itself had a columnar listing of the subunits followed by a 7-point categorical scale. To minimize bias, the directionality of the scale was reversed for one half of the raters. A "1" or "7" on the scale was labeled "Least ____," or "Most ____," and "4" was "average." As with the eliminative method, raters (n = 220) judged each passage on only a single dimension (numbers of raters: concreteness, n = 50; specificity, n = 66; comprehensibility, n = 40; and

interest, $n = 64$). For the statistical analyses, one set of ratings on each dimension was transposed to provide a common directionality. The 7-point ratings of the 60 subunits of "Language" showed a mean correlation of .85 (range = .72–.94) with the number of times that raters retained the subunits in the eliminative method. For the 80 subunits of "Evolution," the mean correlation of the two sets of dimensional ratings was .76 (range = .56–.86).

Measurement of Recall

Undergraduate students in psychology courses were randomly assigned to immediate or delayed reproduction groups. All subjects within a classroom received the same textual excerpt. Learners were advised to "read the textual selection two times at your regular reading rate" and that "sometime in the future you will be tested on the accuracy of your recall." Following the readings, subjects in the delayed groups were dismissed, and the remaining subjects were asked to "write the textual passage as accurately as you possibly can. Recall the exact words and ideas if you can." After reproductions were completed, subjects were requested to avoid discussing the experiment.

Immediate reproductions of "Evolution" were attempted by 58 subjects, and seven-day reproductions were produced by 56 subjects. Reproductions of "Language" were attempted by 61 learners at the immediate interval and 46 learners at the delayed interval. The lower sample sizes at the delayed intervals resulted from absences and from claims that nothing could be remembered. Since the loss in subjects was restricted to the delayed interval, the data for each retention interval were analyzed separately. Due to the attrition, no conclusions may be drawn regarding differential losses in retention as a function of different levels of the semantic dimensions.

The verbatim reproduction of any prose subunit was a rarity, and a subunit was judged to be recalled if the unit was represented in any form in the learner's reproduction. Two trained undergraduate raters independently judged the protocols. Discrepancies in scoring were resolved through conference. If agreement was not reached, a third scorer decided the issue.

For each of the four semantic dimensions, as measured by each of the two methods, the prose subunits of the respective textual passages were rank ordered according to their status on that dimension. Following the rank ordering, the subunits were separated into four groupings ranging from the lowest one fourth on the dimension to the highest one fourth. The learner then received four scores based upon the number of units recalled from the lowest one fourth, the second lowest one fourth, the third lowest one fourth, and the highest one fourth. Since "Language" contained 60 prose subunits, the maximum score at any level was 15. The 80 subunits of "Evolution" were similarly divided into four levels of 20 units.

Results

Table 1 shows mean recall as related to levels of concreteness. Repeated measures analyses of variance were computed separately for the two rating methods and for the two retention intervals. As indicated by the F ratios, each analysis gave evidence that recall was superior for pausal subunits high in concreteness ($ps < .001$). Prose subunits in the upper one fourth in concreteness were recalled approximately three to nine times better than subunits in the lower one fourth in concreteness. The facilitative effects of concreteness are consistent with previous research focusing on the recall of individual words from sets of unrelated sentences (James, 1972), and the overall recall of passages differing in their constituent proportions of abstract or concrete words (Yuille & Paivio, 1969).

In paired-associate learning, higher levels

TABLE 1

Mean Recall of Textual Subunits as a Function of Abstractness–Concreteness

Textual passage	Dimensional ratings	Retention interval	Levels of concreteness				F*	df
			Lowest	Second	Third	Highest		
"Language"	eliminative	immediate	1.28	3.61	6.03	9.30	252.10	3/180
		delayed	.43	1.50	2.43	3.78	36.01	3/135
"Evolution"	eliminative	immediate	3.26	6.03	9.83	10.84	165.84	3/171
		delayed	.95	2.80	5.20	5.54	79.83	3/165
"Language"	categorical	immediate	2.02	2.62	6.57	9.00	276.43	3/180
		delayed	.65	1.02	2.78	3.70	40.03	3/135
"Evolution"	categorical	immediate	2.66	6.95	8.09	12.28	242.56	3/171
		delayed	.73	3.21	3.43	7.11	130.66	3/165

* All $ps < .001$.

TABLE 2

MEAN RECALL OF TEXTUAL SUBUNITS AS A FUNCTION OF SPECIFICITY OF DENOTATION

Textual passage	Dimensional ratings	Retention interval	Levels of specificity				F^*	df
			Lowest	Second	Third	Highest		
"Language"	eliminative	immediate	2.36	3.54	5.52	8.79	153.36	3/180
		delayed	1.07	1.50	2.17	3.41	22.10	3/135
"Evolution"	eliminative	immediate	3.71	4.98	10.24	11.03	187.79	3/171
		delayed	1.95	1.50	4.85	6.20	101.11	3/165
"Language"	categorical	immediate	1.03	4.03	6.18	8.97	217.86	3/180
		delayed	.43	1.61	2.74	3.37	39.91	3/135
"Evolution"	categorical	immediate	3.71	5.36	8.47	12.43	174.90	3/171
		delayed	1.68	2.62	3.41	6.77	87.92	3/165

* All $ps < .001$.

of stimulus specificity are associated with superior remembering (Paivio & Olver, 1964). As shown in Table 2, prose units rated high in specificity also are more likely to be recalled ($ps < .001$). The only numerical reversal to the relationship was not significant in a Newman-Keuls analysis ($p > .05$).

Research on comprehensibility gives indirect support to the expectation that verbal units high in comprehensibility would be more likely to be retained (Carroll, 1971). In the present experiment, as shown in Table 3, higher comprehensibility was associated with superior remembering ($ps < .001$). The only significant reversal of trend occurred between "Evolution's" lowest and second levels as determined by the eliminative method ($ps < .01$). With the categorical method, the lowest and second levels did not differ significantly ($ps > .05$).

Although interest is reputed to be a prime determinant of remembering, a search of the literature uncovered only two studies, and both studies found no relationship between interest and the retention of discourse (Becker, 1963; Klare, Mabry, & Gustafson, 1955). In the present study, as shown in Table 4, higher levels of interest were associated with superior remembering ($ps < .001$). The numerical reversals to the general trend were not significant ($ps > .05$).

To summarize, the four textual dimensions were strongly related to both immediate and delayed recall. Regardless of the method of measurement, prose subunits rated high on a particular dimension were associated with superior recall.

The comparability of the four dimensions in predicting recall suggests the possibility that the dimensions are simply overlapping measures of the same entity. Empirically,

TABLE 3

MEAN RECALL OF TEXTUAL SUBUNITS AS A FUNCTION OF COMPREHENSIBILITY

Textual passage	Dimensional ratings	Retention interval	Levels of comprehensibility				F^*	df
			Lowest	Second	Third	Highest		
"Language"	eliminative	immediate	2.97	5.33	5.21	6.75	65.49	3/180
		delayed	.98	2.15	2.52	2.50	17.87	3/135
"Evolution"	eliminative	immediate	8.14	5.90	7.84	7.81	15.62	3/171
		delayed	4.07	2.64	3.41	4.32	5.79	3/165
"Language"	categorical	immediate	2.77	3.66	5.97	7.82	128.32	3/180
		delayed	.87	1.26	3.17	2.85	31.27	3/135
"Evolution"	categorical	immediate	6.19	6.14	9.03	8.60	46.39	3/171
		delayed	2.27	2.20	4.91	5.11	46.13	3/165

* All $ps < .001$.

TABLE 4

MEAN RECALL OF TEXTUAL SUBUNITS AS A FUNCTION OF INTEREST

Textual passage	Dimensional ratings	Retention interval	Levels of interest				F*	df
			Lowest	Second	Third	Highest		
"Language"	eliminative	immediate	3.97	5.26	5.16	5.82	17.93	3/180
		delayed	1.13	2.24	2.37	2.41	12.15	3/135
"Evolution"	eliminative	immediate	5.62	6.50	7.88	9.97	64.20	3/171
		delayed	3.11	2.57	3.71	5.09	30.93	3/165
"Language"	categorical	immediate	3.84	4.02	6.10	6.26	39.50	3/180
		delayed	1.07	1.74	2.48	2.87	21.51	3/135
"Evolution"	categorical	immediate	3.29	7.72	7.83	11.12	145.09	3/171
		delayed	1.43	3.64	3.36	6.05	64.09	3/165

* All $ps < .001$.

each textual dimension was significantly correlated with at least one other textual dimension.[4] On a theoretical basis, too, the textual dimensions appear to be closely interrelated. Whether a subunit was judged to be interesting, for example, might depend upon the ease of comprehension, or perhaps directly upon concreteness or specificity.

In assessments of the independence of the textual dimensions, stepwise multiple regression analyses and factor analyses were computed. For each textual passage, the dependent variable in the regression analyses was the number of recalls of each textual subunit. The potential predictors included the eight measures of the textual dimensions, the number of words in each subunit, and serial position. Entry into the regression equation was permitted only when the new predictor accounted for a significant increment of the residual variance.

Analysis of the immediate recalls of the subunits of "Language" showed three variables accounting for independent portions of the recall variance. The order of entry into the equation, and the results of tests to determine entry were concreteness, eliminative scaling ($F = 109.14$, $df = 1/58$, $p < .001$); comprehensibility, eliminative scaling ($F = 9.15$, $df = 1/57$, $p < .01$); and comprehensibility, categorical scaling ($F = 4.70$, $df = 1/56$, $p < .05$). The optimal set of three predictors showed a multiple correlation of .85 with immediate recall. In the delayed recall of "Language," an optimal set of three variables showed a multiple correlation of .74 with the criterion variable: concreteness, eliminative scaling ($F = 41.38$, $df = 1/58$, $p < .001$); serial position ($F = 9.37$, $df = 1/57$, $p < .01$); and interest, categorical scaling ($F = 5.60$, $df = 1/56$, $p < .05$). The immediate recall of "Evolution" showed a multiple correlation of .79 with a set of five independent predictors: specificity, categorical scaling ($F = 49.51$, $df = 1/78$, $p < .001$); comprehensibility, categorical scaling ($F = 21.62$, $df = 1/77$, $p < .001$); specificity, eliminative scaling ($F = 4.97$, $df = 1/76$, $p < .05$); comprehensibility, eliminative scaling ($F = 6.01$, $df = 1/75$, $p < .05$); and interest, categorical scaling ($F = 6.31$, $df = 1/74$, $p < .05$). A set of six variables showed a multiple correlation of .75 with the delayed recall of "Evolution"; concreteness, categorical scaling ($F = 30.53$, $df = 1/78$, $p < .001$); specificity, eliminative scaling ($F = 6.28$, $df = 1/77$, $p < .05$); serial position, ($F = 8.13$, $df = 1/76$, $p < .01$); comprehensibility, categorical scaling ($F = 8.61$, $df = 1/75$, $p < .01$); comprehensibility, eliminative scaling ($F = 10.40$, $df = 1/74$, $p < .01$); and interest, categorical scaling ($F = 5.07$, $df = 1/74$, $p < .05$).

In summary, for each textual passage and each retention interval, the regression analyses provided evidence that at least two textual dimensions were independent predictors of recall. The particular variables entering into the prediction equations differed somewhat in the various analyses, but each

[4] Tables showing the intercorrelations among variables are available from the author.

textual dimension emerged as an independent predictor in at least two analyses. In addition, serial position did not predict immediate recall, but primacy effects existed in delayed retention.

Factor analyses provided additional information on the independence of the various textual dimensions. Principal-components solutions, with communalities estimated as unity, were followed by orthogonal rotation. With the criterion for rotation set at eigenvalues greater than unity, three factors were rotated in each of the prose passages. As defined by factor loadings of .30 or higher, the variables (and loadings) of Factor 1 of "Language" were concreteness, categorical scaling (.95); concreteness, eliminative scaling (.95); comprehensibility, categorical scaling (.87); comprehensibility, eliminative scaling (.82); specificity, categorical scaling (.57); and specificity, eliminative scaling (.49). Factor 2 was defined by interest, categorical scaling (.87); interest, eliminative scaling (.83); number of words (.77); specificity, eliminative scaling (.63); specificity, categorical scaling (.61); and concreteness, eliminative scaling (.36). The third factor was composed of serial position (.90), and specificity, eliminative scaling (.35). With immediate recall in the analysis, the variable loaded on Factor 1 (.78) and Factor 2 (.37). Delayed recall was included in Factor 1 (.60), Factor 2 (.42), and Factor 3 (−.42).

The first factor of "Evolution" was defined by specificity, categorical scaling (.91); specificity, eliminative scaling (.87); concreteness, eliminative scaling (.81); concreteness, categorical scaling (.74); interest, categorical scaling (.76); interest, eliminative scaling (.66); and number of words (.36). Factor 2 included comprehensibility, categorical scaling (.92); comprehensibility, eliminative scaling (.88); concreteness, categorical scaling (.45); concreteness, eliminative scaling (.40); interest, categorical scaling (−.36), and number of words (−.80). The third factor consisted of serial position (.88); interest, eliminative scaling (.39); and concreteness, categorical scaling (−.35). Immediate recall loaded only on Factor 1 (.76), while delayed recall had loadings on Factor 1 (.51) and Factor 3 (−.58).

Overall, the factor analyses provided corroborative evidence of some independence among the textual dimensions. Three factors were required to account for variations among the textual dimensions. Further, the entry of recall variables into more than one factorial cluster of textual dimensions suggests that orthogonal dimensions of prose can account for independent portions of the recall variance.

Discussion

Recall was strongly related to the semantic dimensions of abstractness–concreteness, specificity of denotation, comprehensibility, and interest. Unlike previous research with the semantic dimensions, the psycholinguistic segments were partitioned in consecutive serial order, without omission, from a larger passage of intact textual prose. The successful predictions of the patternings of textual recall thus were made for all subunits within a prose passage. Attesting to the generality of the findings, the relationships were evident with two textual passages, with two methods of measuring the dimensions, and at two retention intervals. The strength of the relationships between recall and each of the textual dimensions also is evident from the fact that the prose subunits were not selected to differ maximally on any one of the textual dimensions, and each dimension was compared with the same sets of recall data.

The similarity in factorial solutions for the two textual passages suggests some invariance in the factorial composition of textual prose. Complexities in structure, however, also are evident from the emergence of multiple factors. If additional textual dimensions had been entered into the factor analyses, additional factors probably would have emerged. Carroll's (1960) factor analysis of variations between passages, for example, led to a solution with six interpretable factors, and a comparable complexity probably exists among subunits within a single passage.

Complexities in the semantic structure of prose suggest the possibility of different

influences on recall. In the free recall of nouns, for example, adjectival modification led to an inferior remembering of nouns (Cofer, Segal, Stein, & Walker, 1969). For prose, however, adjectival modification of nouns neither aided nor hindered recall, whereas the specificity level of nouns per se was a determinant of remembering (August, Proctor, Hynes, & Johnson, 1974). In the present study, then, the various determinants of specificity presumably did not exert a unitary influence on recall. Even so, denotative specificity was strongly associated with remembering, and this finding suggests the appropriateness of reexamining the data relevant to the widely quoted principle that generalizations are remembered better than specific facts (e.g., McDougall, 1958).

The dimensional ratings of the textual segments were inversely related to the patterning of selective omissions. Although such evidence is consistent with Zangwill's (1972) contention that the learning of prose is an abstractive process, the data do not argue against a reconstructive theory of remembering. Cofer (1973) contrasts the processes of reconstruction and abstraction as alternative explanations of remembering, but the processes are not mutually exclusive. Separate ideational units in prose do become amalgamated in remembering (e.g., Bartlett, 1932), but the occurrence of selective omissions also demonstrates that certain types of ideational units are more likely to survive in recall.

Since various textual dimensions were independent predictors of recall, the process of abstraction appears to be influenced by a number of textual characteristics. Further, since selective omissions in immediate recall are evident even when equivalent learning times are allocated to each subunit (Johnson, 1970, Experiment III), the abstractive process does not appear dependent upon differential allocations of processing time in learning or upon differential rehearsals after apprehension. The occurrence of abstraction, however, could stem from differential processing of the various prose units at the time of learning. Such selective processing might be derived from learner strategies or from existing knowledge structures, and an examination of the relation between textual ratings and prereading cloze performances would provide relevant evidence on this issue. Since selective omissions were evident at immediate and delayed retention intervals, the events occurring during the retention interval do not appear implicated in the abstractive process. Another possibility, though, is that abstraction is determined by conditions existing at the time of recall. With sufficient prompts for recall, for example, the data might show that omitted units are represented in memory even though such units ordinarily are not recalled.

At the applicational level, the modification of textual content toward greater comprehensibility, interest, concreteness, or specificity presumably would result in increases in recall. Since dimensional variations might be correlated with differences in information load or predictability, appropriate empirical tests would require the control of rates of information transmission (Rubenstein & Aborn, 1958).

REFERENCES

August, G. D., Proctor, D. L., Hynes, K. P., & Johnson, R. E. Recall of prose as a function of denotative specificity. Paper presented at the American Psychological Association, New Orleans, September, 1974.

Bartlett, F. C. *Remembering*. London: Cambridge University Press, 1932.

Becker, S. L. *The relationships of interest and attention to retention and attitude change*. (Doctoral dissertation, University of Iowa) Iowa City, Iowa: University Microfilms, 1963, No. 64-4898.

Carroll, J. B. Vectors of prose style, In T. A. Sebeok (Ed.), *Style in language*. Cambridge, Mass.: MIT Press, 1960.

Carroll, J. B. *Language and thought*. Englewood Cliffs, N.J.: Prentice-Hall, 1964.

Carroll, J. B. *Learning from verbal discourse in educational media: A review of the literature*. (Research Bulletin-71-61) Princeton, N.J.: Educational Testing Service, 1971.

Cofer, C. N. Constructive processes in memory. *American Scientist*, 1973, **61**, 537–543.

Cofer, C. N., Segal, E., Stein, J., & Walker, H. Studies on free recall of nouns following presentation under adjectival modification. *Journal of Experimental Psychology*, 1969, **79**, 254–264.

Gomulicka, B. R. Recall as an abstractive process. *Acta Psychologica*, 1956, **12**, 77–94.

James, C. T. Theme and imagery in the recall of active and passive sentences. *Journal of Verbal*

Learning and Verbal Behavior, 1972, **11,** 205–211.

Johnson, R. E. The retention of qualitative changes in learning. *Journal of Verbal Learning and Verbal Behavior,* 1962, **1,** 218–223.

Johnson, R. E. Recall of prose as a function of the structural importance of the linguistic units. *Journal of Verbal Learning and Verbal Behavior,* 1970, **9,** 12–20.

Johnson, R. E. Meaningfulness and the recall of textual prose. *American Educational Research Journal,* 1973, **10,** 49–58.

Klare, G. R., Mabry, J. E., & Gustafson, L. M. The relationship of human interest to immediate retention and to acceptability of technical material. *Journal of Applied Psychology,* 1955, **39,** 92–95.

McDougall, W. P. Differential retention of course outcomes in educational psychology. *Journal of Educational Psychology,* 1958, **49,** 53–60.

Meyer, B. J. F., & McConkie, G. W. What is recalled after hearing a passage? *Journal of Educational Psychology,* 1973, **65,** 109–117.

Munn, N. L. *Psychology.* Boston: Houghton Mifflin, 1946.

Paivio, A., & Olver, M. Denotative-generality, imagery, and meaningfulness in paired-associate learning of nouns. *Psychonomic Science,* 1964, **1,** 183–184.

Rubenstein, H., & Aborn, M. Learning, prediction, and readability. *Journal of Applied Psychology,* 1958, **42,** 28–32.

Wickens, D. D., & Meyer, D. R. *Psychology.* New York: Dryden, 1955.

Yuille, J. C., & Paivio, A. Abstractness and recall of connected discourse. *Journal of Experimental Psychology,* 1969, **82,** 467–471.

Zangwill, O. L. Remembering revisited. *Quarterly Journal of Experimental Psychology,* 1972, **24,** 123–138.

(Received November 27, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 780–786

ENCODING SPEED, REHEARSAL, AND PROBED RECALL OF DYSLEXIC BOYS[1]

CARL SPRING[2] AND CAROLYN CAPPS

University of California, Davis

A model attributing poor recall to slow speech–motor encoding is described. The following results, predicted from the model, were obtained with poor readers of normal intelligence ($n = 24$) and normal readers ($n = 24$) from ages 7 to 13: (*a*) Poor readers named visually presented, nonword stimuli more slowly than normal readers; (*b*) fewer poor readers than normal readers employed a cumulative rehearsal strategy during a probed-recall task; (*c*) the use of cumulative rehearsal was significantly related to naming speed; (*d*) the performance of poor readers was inferior to that of normal readers for all but the most recently presented items of the probe-recall task; and (*e*) naming speed and use of cumulative rehearsal accounted for 91% of the true variance of early and middle serial positions of the probed recall task.

Certain children, although they test from normal to high on intelligence and have had adequate instruction, learn to read and spell only with extraordinary difficulty, if at all. In this report, poor readers of this type are called dyslexic children. Although dyslexic children generally perform well on several subtests of the Wechsler Intelligence Scale for Children (WISC), investigators have consistently found that their scores are low on the Digit Span subtest (Klasen, 1972). In addition to the Digit Span subtest, dyslexic children are deficient on a variety of other serial memory tests using visual and auditory input (Doehring, 1968). The present study tested a model attributing the impaired recall of dyslexic children to slow speech–motor encoding.

The hypothesized relation of memory and encoding speed is based on a two-storage memory model (Atkinson & Shiffrin, 1968). As an item is presented, it is covertly encoded to a speech–motor response and briefly stored in a highly transient short-term storage. Items in short-term storage are rapidly forgotten due to displacement by subsequent items. If items in short-term storage are rehearsed, they may be transferred to a more durable long-term storage. Increasing the presentation rate, however, decreases the time between presentations that is available for rehearsal, thereby reducing the probability of recalling all but the most recently presented items (Ellis & Hope, 1968; Glanzer & Cunitz, 1966). Recall of the two or three most recently presented items is not affected by increasing the rate of presentation because of the high probability that these items are still available in short-term storage.

Unusually slow speech–motor encoding should have an effect similar to a fast rate of presentation. For individuals who are unusually slow speech–motor encoders, little time is available for rehearsal between item presentations. This should reduce the probability of recalling all but the most recently presented items; therefore, the performance of dyslexic children, who are hypothesized to be slow encoders, should be inferior to that of normal children for all but the most recently presented items. In addition, there should be a strong correlation between recall and encoding speed.

[1] This research was partially supported by Grant MH-20583 from the National Institute of Mental Health. The authors wish to acknowledge the cooperation of the San Juan Joint Unified School District.

[2] Requests for reprints should be sent to Carl Spring, Department of Education, College of Letters and Science, University of California, Davis, California 95616.

Speech–motor encoding speed was measured, in the present study, by asking subjects to name digits, color patches, and pictures as rapidly as possible. It is well known that verbal stimuli such as digits, letters, and words are named faster than concrete stimuli such as color patches and pictures. In a previous study, the difference between encoding speeds of dyslexic and normal children was larger for digits (verbal stimulus) than for colors and pictures (concrete stimuli).[3] A similar interaction was expected in the present study.

Memory was tested with a probe-recall task. A series of cards were briefly exposed one at a time and placed face down before the subject. Each card contained a digit. A probe digit was then presented, and the subject's task was to point to the face-down card with the matching digit.

We were able, using the probe task, to observe individual differences in visual scanning patterns. Atkinson, Hansen, and Bernbach (1964) timed responses of preschool children on a probe-recall task. They found that response latency increased linearly from the most recent to the earliest probed positions, and they interpreted this as evidence that subjects counted *backwards* from the most recent item, using interitem associations to determine how far back the probed item occurred. In a pilot study with young adults, we obtained results suggesting a different strategy. Following presentation of the probe, eye movements were observed. These older subjects typically scanned *forward* from the first card, stopping on the card of their choice. It seems reasonable that direction of scanning should differ at different developmental levels. This difference probably reflects the employment of a cumulative rehearsal strategy in which an attempt is made to recall previously presented items in their correct order each time a new item is presented. The effectiveness of cumulative rehearsal, of course, depends on locating the probed item by visually scanning forward from the first card while reciting the items in the rehearsed order. Because many dyslexic children were expected to encode too slowly to permit cumulative rehearsal, we expected that fewer dyslexic than normal children would employ forward scanning. In addition, a strong correlation between encoding speed and evidence of forward scanning was predicted.

[3] Unpublished study entitled "Encoding Speed and Memory Span in Dyslexic Children," 1973.

Method

Subjects

Two groups of boys, ranging from 7.6- to 13.4-years-old, were selected from public elementary schools. The groups were closely matched on age and approximately matched on father's occupation as indexed by the Hollingshead Scale of Social Position. One group ($n = 24$), designated dyslexic, was composed of poor readers selected from special classes. Boys in this group were tested on the Wide Range Achievement Test. They scored, on the average, 2.3 years below expected grade level in reading, 2.8 years below grade level in spelling, and 1.9 years below grade level in arithmetic. Recent WISC IQ scores were available for these boys. Their mean full-scale IQ was 102, and scores ranged from 90 to 115. A second group ($n = 24$), designated normal, was selected from boys in regular classrooms whose reading abilities were close to their expected grade levels. Depending on grade, their reading levels were measured by the Cooperative Primary Reading Test, Stanford Achievement Test, or California Test of Basic Skills. The average reading score of normal readers was .2 years above expected grade level. The intelligence of normal readers was not determined.

Tasks

In the probe-recall task, eight cards were exposed one at a time and placed face down, from subject's left to right. The cards, presented at a 1.5-second rate, contained the Digits 1 through 8 in random order. A probe digit was then presented, and subject's task was to point to the card with the matching digit. If the first choice was incorrect, a second choice was given. Twenty-four trials were given, preceded by 2 practice trials. This allowed each position to be probed three times. Within blocks of 8 consecutive trials, all 8 positions were probed once in random order.

Digit naming was tested by requiring subjects to name 50 randomly sequenced digits, typed in a single row, as fast as possible. Naming was from left to right. Two-syllable digits (0 and 7) were excluded. Time and errors were recorded. Two forms were developed for the digit-naming test.

Color naming was tested by requiring subjects to name 30 color patches as fast as possible. The color patches were ¾-inch squares arrayed on an 8 × 10 inch plate in 5 rows and 6 columns. Naming was from left to right and top to bottom. The

following colors were sequenced randomly, except that no two adjacent color patches were the same: red, orange, yellow, blue, green, black, and brown. Time and errors were recorded.

Picture naming was tested by requiring subjects to name 25 tinted line-drawings as fast as possible. The drawings were on 2-inch squares arrayed on a 12 × 12 inch plate in 5 rows and 5 columns. Naming was from left to right and top to bottom. The following pictures were used: wagon, pear, duck, umbrella, girl, gun, nurse, moon, fish, orange, leaf, cat, flag, monkey, tree, pig, ball, dog, house, apple, turtle, lamp, kite, king, and star. Time and errors were recorded.

Procedure

Subjects were tested individually on identical tasks in a fixed order. Two testers were used. Half of the boys tested by each tester were normal readers, and half were poor readers. Testers did not know which boys were poor readers. Naming tasks were given in the following order: digits (Form A), colors, pictures, digits (Form B). After each naming task, six trials of the probe-recall task were given.

Results and Discussion

Naming Tests

Mean error rates of dyslexic and normal subjects were less than .02 for each type of stimulus. Although dyslexic boys tended to make more errors than normal boys, differences were small and not significant.

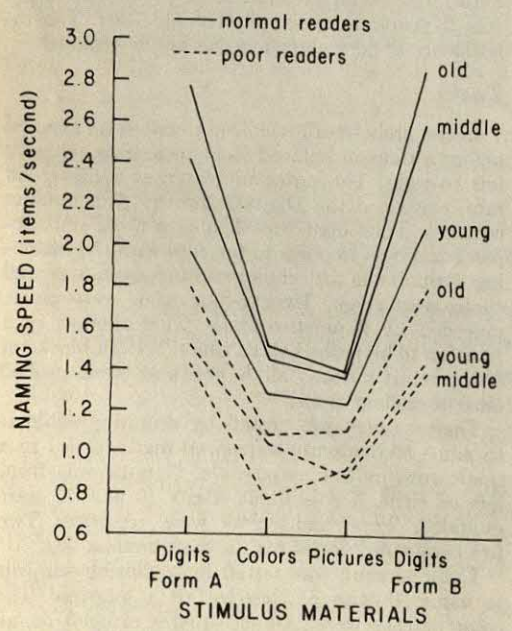


Figure 1. Mean naming speeds of dyslexic and normal boys for digits, colors, and pictures.

Mean naming speeds, expressed as items per second, are shown in Figure 1. These data were analyzed with a 2 × 3 × 4 repeated measures analysis of variance. The first factor was reading ability, with two levels: dyslexic and normal. The second factor was age, with three levels; young (91–120 months), middle (121–144 months), and old (145–161 months). The third factor was stimulus, with four repeated measures: digits (Form A), colors, pictures, and digits (Form B).

All main effects were significant. Dyslexic boys were slower than normal boys ($F = 63.8$, $df = 1/42$, $p < .001$). Thus, a major hypothesis was confirmed. In addition, younger children were slower than older children ($F = 7.8$, $df = 2/42$, $p < .005$), and performance on colors and pictures (concrete stimuli) was slower than performance on digits (verbal stimulus; $F = 197.5$, $df = 3/126, p < .001$).

A significant Age × Stimulus interaction was found ($F = 3.4$, $df = 6/126, p < .005$). This interaction was due to a larger difference between younger and older subjects on digit-naming speed than on speed of naming colors and pictures. The interaction may be due to the possibility that children, as they advance in school, gradually accumulate more practice in speech–motor encoding of verbal stimuli than concrete stimuli. This would lead to a gradual divergence of encoding speeds for verbal and concrete stimuli as children grow older.

As expected, a significant Ability × Stimulus interaction was also found ($F = 18.6$, $df = 3/126$, $p < .001$). This interaction was due to a larger difference between dyslexic and normal subjects on digit-naming speed than on speed of naming colors and pictures. The interaction is congruent with the common clinical observation that dyslexic children are specifically impaired on tasks requiring perception of verbal material, while they evidence no dramatic inability to function in an environment of concrete stimuli. This Ability × Stimulus interaction may be explained in different ways. First, the interaction may be

due to the possibility that dyslexic children practice speech–motor encoding of verbal material, relative to concrete material, less than normal children. This may be due to a lack of motivation to complete school assigments. The interaction may be explained differently, however. Dyslexic and normal children may practice speech–motor encoding equally, with both groups accumulating more practice for verbal than concrete material. But due to a neurological deficit, encoding speed may be enhanced by practice less for poor readers than for normal children. This combination of experiential and neurological factors would also produce an Ability × Stimulus interaction.

Memory Test

Following presentation of the probe on each trial of the memory test, eye movements were observed. Subjects who typically scanned the cards from left to right, finally stopping on the card of their choice, were classified as scanners. All others were classified as nonscanners. As expected, the incidence of nonscanners was significantly greater among dyslexic boys. There was only 1 nonscanner in the normal group, while 11 dyslexic boys were nonscanners ($\chi^2 = 9.0$, $p < .01$, two-tailed). This difference indicates that significantly fewer dyslexic boys employed cumulative rehearsal than normal boys. Evidence supporting this interpretation will be presented later in this section.

Three serial-position curves, showing percentages of correct first-choice responses for each position, are shown in Figure 2. The performance of normal boys is shown by a single curve which includes the data from the lone nonscanner. The performance of dyslexic boys is shown by two curves: one for scanners and another for nonscanners. All subjects were more likely to recall late items than middle items (recency effect). In addition, normals and dyslexic scanners were more likely to recall early items than middle items (primacy effect). A primacy effect, however, was not obtained for dyslexic nonscanners. Thus, only two of the three curves have the bow shape that

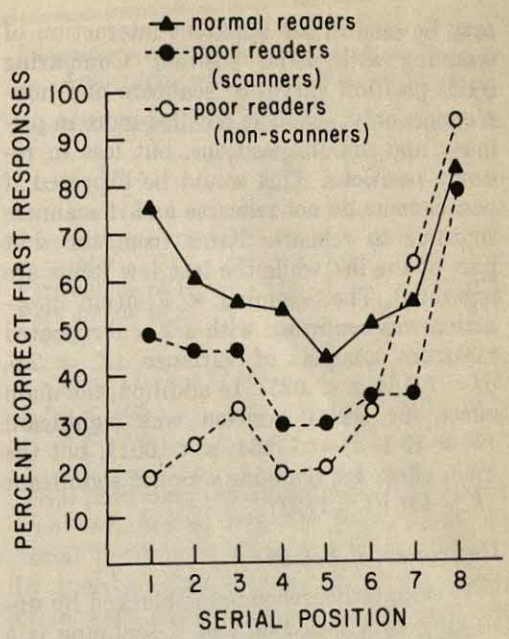


FIGURE 2. Percentages of correct first-choice responses at each serial position for dyslexic boys (scanners and nonscanners), and normal boys.

is characteristic of serial-position recall curves.

As hypothesized, the performance of dyslexic boys (scanners and nonscanners combined) was inferior to that of normal boys in all but the last few positions. This Ability × Position interaction was confirmed with a 2 × 8 repeated measures analysis of variance ($F = 2.7$, $df = 7/332$, $p < .01$). In addition, significant main effects were obtained for reading ability ($F = 26.6$, $df = 1/46$, $p < .001$) and for serial position ($F = 13.7$, $df = 7/332$, $p < .001$).

As noted above, the higher incidence of nonscanners in the dyslexic group indicates that fewer dyslexic boys employed cumulative rehearsal than normal boys. This interpretation is supported by the failure to obtain a primacy effect for dyslexic nonscanners, while clear primacy effects were obtained for normals and dyslexic scanners. If, as is commonly believed, a primacy effect is due to cumulative rehearsal of initial items (Rundus & Atkinson, 1970; Tulving, 1968), these data indicate that nonscanners do not cumulatively rehearse.

Further evidence for this interpretation

may be seen in the apparent interaction of scanning with serial position. Comparing serial-position curves of scanners and nonscanners only, scanners recalled more in primacy and middle positions, but less in recency positions. This would be expected if nonscanners do not rehearse and if scanners continue to rehearse items from the first part of the list while the last few items are presented. The Scanning × Position interaction was confirmed with a 2 × 8 repeated measures analysis of variance ($F = 2.6$, $df = 7/154$, $p < .025$). In addition, the main effect for serial position was significant ($F = 12.1$, $df = 7/154$, $p < .001$), but the main effect for scanning was not significant ($F = 1.0$, $df = 1/22$).

Correlates of Scanning

If cumulative rehearsal is blocked by unusually slow encoding and if scanning is a valid measure of rehearsal, there should be a significant point-biserial correlation between scanning and naming speed. This was tested, for poor readers only, with two correlations that differed by their use of verbal or concrete stimuli to measure naming speed. For the correlation involving verbal stimuli, digit-naming speed was computed by averaging scores from alternate forms of the digit-naming test. For the correlation involving concrete stimuli, color–picture-naming speed was computed by averaging scores from the color and picture tests. Scanning was coded as a two-value categorical variable. Contrary to expectation, the correlation of scanning and digit-naming speed was not significant ($r = .18$, $F < 1$, $df = 1/22$). The correlation of scanning and color–picture-naming speed, however, was highly significant ($r = .57$, $F = 10.6$, $df = 1/22$, $p < .005$). To illustrate the latter relation, naming speeds of poor readers are ranked in Table 1. Beside each entry is a scanning code. It may be seen that a naming speed of one item per second separates scanners from nonscanners with 88% accuracy. The point-biserial correlation of scanning with age, also computed for dyslexic boys, was not significant ($r = .23$, $F = 1.2$, $df = 1/22$).

TABLE 1

SCANNERS AND NONSCANNERS RANKED BY NAMING SPEED

Color–picture-naming speed[a]	Scanning code[b]	Color–picture-naming speed[a]	Scanning code[b]
.67	0	1.01	1
.68	0	1.02	1
.71	0	1.03	1
.75	0	1.07	1
.76	0	1.08	1
.78	1	1.14	1
.80	1	1.14	1
.83	0	1.17	1
.83	0	1.17	1
.85	0	1.36	0
.89	0	1.40	1
.96	0	1.41	1

Note. This refers to dyslexic boys only.
[a] Items per second.
[b] Nonscanners coded 0; scanners coded 1.

These results are puzzling. Why should digit rehearsal be significantly related to encoding speed for colors and pictures but not to encoding speed for digits? The correlations suggest that children do not rehearse digits, regardless of their encoding ability for digits, until they are able to rehearse colors and pictures. Since encoding speed develops more slowly for concrete items (colors and pictures) than for verbal items (digits), the emergence of cumulative rehearsal in a child's behavior, regardless of the type of item to be rehearsed, appears to be delayed until he is able to encode concrete items at speeds above some minimum value. Judging from the present data, this minimum value is about one item per second.

Correlates of Long-Term Storage

It was found that recall probabilities of normals and dyslexics (scanners and nonscanners combined) differed in all but the last two positions. According to a two-storage model, this pattern indicates a long-term-storage difference. It was hypothesized that this long-term-storage difference is caused by a rehearsal deficit. From results already presented, it appears that slow speech–motor encoding may limit cumulative rehearsal or block it completely. In the first case, recall should be related to digit-naming speed. In the second case, it

should be related to scanning. To determine the relative strengths of these relationships, the multiple correlation of long-term storage with digit-naming speed and scanning was computed. Scanning was coded as a two-value categorical variable. Long-term storage was determined for each subject by totaling the number of correct first-choice responses in Serial Positions 1 through 6. The coefficient-alpha reliability of long-term storage was an r_{xx} of .65, indicating a theoretical maximum correlation of long-term storage with other variables of .80 ($r_{xx}^{1/2}$). The obtained multiple correlation of .77 was very near the theoretical upper limit. Putting it another way, digit-naming speed and scanning together accounted for 91% of the true variance of long-term storage (100 R^2/r_{xx}). The relative influence of these variables was examined with a commonality analysis (Kerlinger & Pedhazur, 1973). The results, shown in Table 2, indicate that the variables *commonly* accounted for a significant portion of the variance of long-term storage. Moreover, each variable *uniquely* accounted for an additional significant portion of long-term storage variance. The portion unique to digit-naming speed, however, was much larger than the portion unique to scanning.

TABLE 2

PROPORTION OF VARIANCE OF LONG-TERM STORAGE ASSOCIATED WITH DIGIT-NAMING SPEED AND SCANNING

Source	1. Digit-naming speed	2. Scanning	F^a
Part common to 1 and 2	.25	.25	26.8**
Part unique to 1	.28		30.0**
Part unique to 2		.06	6.4*
Total	.53	.31	

[a] $df = 1/44$.
* $p < .025$.
** $p < .001$.

CONCLUSIONS

Data indicated the following: (*a*) Dyslexics were slower than normals and younger boys were slower than older boys on naming speed tests; (*b*) these differences were larger for digits (a verbal stimulus) than for colors and pictures (concrete stimuli); (*c*) subjects in both groups named digits faster than colors and pictures; (*d*) almost all of the normals, but barely half of the dyslexics, employed left-to-right visual scanning during the probe task; (*e*) there were primacy effects in the serial-position curves of normals and dyslexic scanners but not of dyslexic nonscanners; (*f*) speed of naming colors and pictures discriminated between dyslexic scanners and nonscanners, with lower speeds for nonscanners; (*g*) probability of correct recall was greater for normals than dyslexics (scanners and nonscanners) in all but the most recent two serial positions; (*h*) almost all (91%) of the true variance of memory performance in Serial Positions 1 through 6 was predicted by speed of digit naming and employment of forward scanning; and (*i*) digit naming was a more powerful predictor of memory performance than forward scanning.

These data are interpreted as follows. The interaction of reading ability with serial position indicates that short-term storage is equally robust for dyslexics and normals but that long-term storage is impaired for dyslexics. As hypothesized, impaired long-term storage is strongly related to slow speech–motor encoding, presumably because slow encoding preempts time that would otherwise be available for rehearsal. It appears that slow encoding acts in two ways. It may merely limit rehearsal or block it completely. When digits are used in the memory task, the time available for rehearsal is limited by the subject's encoding speed for digits. Whether rehearsal will even be attempted, however, is predicted best by speed of naming concrete stimuli such as colors and pictures, perhaps because rehearsal is a habit that is not acquired until children are able to encode *any* type of labeled stimulus at a speed in excess of a threshold value. The present data indicate a threshold value of about one item per second.

This analysis probably does not apply equally well to all poor readers. Poor readers with IQs above 90 were selected for the present study. There are many more

school children with IQs *below* 90 who also have serious reading problems. Unlike the children studied in the present investigation, however, many low-IQ children perform in the normal range on simple learning tasks such as serial memory, although they are less skilled on more complex tasks requiring abstract reasoning (Jensen, 1973). Their low IQs apparently reflect the fact that intelligence tests are heavily loaded with items requiring abstract reasoning. Dyslexic children, on the other hand, have WISC subtest scores indicating impaired serial memory but normal performance on tasks requiring abstract reasoning. The conclusions and interpretations presented in this report are limited to the latter group.

REFERENCES

Atkinson, R. C., Hansen, D. N., & Bernbach, H. S. Short-term memory with young children. *Psychonomic Science*, 1964, **1**, 255–256.

Atkinson, R. C., & Shiffrin, R. M. Human memory: A proposed system and its control processes. In K. W. Spence and J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory*. Vol. 2. New York: Academic Press, 1968.

Doehring, D. G. *Patterns of impairment in specific reading disability*. Bloomington: Indiana University Press, 1968.

Ellis, N. R., & Hope, R. H. Memory processes and the serial position curve. *Journal of Experimental Psychology*, 1968, **77**, 613–619.

Glanzer, M., & Cunitz, A. R. Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, 1966, **5**, 351–360.

Jensen, A. R. Level I and level II abilities in three ethnic groups. *American Educational Research Journal*, 1973, **10**, 263–276.

Kerlinger, F. N., & Pedhazur, E. J. *Multiple regression in behavioral research*. New York: Holt, Rinehart & Winston, 1973.

Klasen, E. *The syndrome of specific dyslexia*. Baltimore, Md.: University Park Press, 1972.

Rundus, D., & Atkinson, R. C. Rehearsal processes in free recall: A procedure for direct observation. *Journal of Verbal Learning and Verbal Behavior*, 1970, **9**, 99–105.

Tulving, E. Theoretical issues in free recall. In T. R. Dixon and D. L. Horton (Eds.), *Verbal behavior and general behavior theory*. Englewood Cliffs, N.J.: Prentice-Hall, 1968.

(Received December 4, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 5, 787-792

EFFECTS OF LEARNING PROSE WITH PART VERSUS WHOLE PRESENTATIONS OF INSTRUCTIONAL OBJECTIVES[1]

ROBERT KAPLAN[2]

Bell Laboratories, New Brunswick, New Jersey

The purpose of the present study was to determine the effects of part versus whole presentations of objectives and text upon intentional and incidental prose learning. These effects were examined for three passage lengths with specific or general objectives and with two densities of objectives. The results showed that presentation of objectives generally facilitated intentional learning with either increases or no reduction in incidental learning. The part presentation resulted in greater likelihood of learning intentional items than the whole presentation with no significant loss of incidental learning. Neither specificity nor density of objectives was found to influence performance. More inspection time was used for part presentations, longer passages, specific objectives and larger densities.

Providing instructional objectives to students prior to text as directions to learn has been shown to enhance both intentional (objective-relevant) and incidental (non-objective-relevant) learning (Rothkopf & Kaplan, 1972). Further, Kaplan and Rothkopf (1974) found that when objectives were inspected as a whole, the likelihood of learning objective-relevant material was greater with (*a*) shorter passages, (*b*) fewer objectives, and (*c*) more specifically phrased objectives. However, instructional materials are seldom only a few pages in length and contain a small proportion of objective-relevant material. Further, the desirability of short passages and few objectives conflicts with the desirability of specifically phrased objectives. That is, a greater number of specific objectives is required to cover the same instructional points than the more generally phrased objectives. Therefore, the type of instructional material that students are usually required to learn is at variance with the best use of objectives suggested in the Kaplan and Rothkopf (1974) study. The problem is one of finding a way to provide objectives such that the likelihood of learning remains high for large texts containing a large quantity of objective-relevant information. One technique that could resolve this problem is to distribute partial lists of objectives among corresponding text segments. This part presentation method would permit the learner to consider short text segments with relatively few specifically phrased objectives. There is some evidence indicating that this general procedure can enhance learning when questions are distributed throughout text segments (Frase, 1967; Rothkopf & Bisbicos, 1967). However, these studies did not investigate the whole presentation of questions or objectives (all objectives presented prior to the complete text) as in the Kaplan and Rothkopf (1974) study. Papay (1971) did investigate part versus whole presentations of objectives and text. He reported "inconclusive" results for intentional learning with respect to part versus whole presentations, and no comparisons were made for non-objective-relevant learning. In addition, Papay's finding was based on 28 objectives for a 3,500-word text. The Kaplan and Rothkopf (1974) findings for whole presentation were based on a larger proportion of objectives to text sentences (e.g., from 22,

[1] The author is indebted to E. M. Burgin for her valuable assistance in preparing materials and analyzing data.

[2] Requests for reprints should be sent to Robert Kaplan, Bell Laboratories, CPI L101, P.O. Box 2020, New Brunswick, New Jersey 08903.

34, 41, or 48 objectives for a 56-sentence text to 68, 104, 124, or 146 objectives for a 169-sentence text).

The purpose of the present study was to determine if performance increments found with whole presentations of objectives prior to text (Kaplan & Rothkopf, 1974) could be maintained when objectives were distributed among text segments. In addition, the effects upon learning of passage length, specificity of stating objectives, and density (proportion of objective-relevant text sentences to total text sentences) were investigated with part and whole presentations of objectives and text.

Method

Materials

Passages. Three basic passages used in previous studies (Kaplan & Rothkopf, 1974; Rothkopf & Kaplan, 1972) were used in the present study.[3] The passages were 60, 54, and 55 sentences in length ($\bar{X} = 56$ sentences). These individually presented passages were designated Passages 1, 2, and 3, and they comprised the Length 56 condition. The Length 56 passages were then paired to form a Length 113 condition. Three pairs of Length 56 passage combinations (2,1; 2,3; and 3,1), with a mean of 113 sentences, were used. Finally, a Length 169 condition was developed by combining all three Length 56 passages in two different sequential orders (passage combinations 2,1,3 and 2,3,1). The particular passage combination sequences used for Length 113 and 169 were selected for logical flow of information.

Objectives. The objectives were the same as those used in the two studies previously mentioned. Objectives were defined as directions to the subject to learn specific instructional points. Two types of objectives were prepared: specific and general. A specific objective was one that required the subject to learn one passage sentence. Thus, the number of specific objectives equaled the number of objective-relevant passage sentences. The match between objectives and passage sentences was empirically determined in a pre-experimental study (see Rothkopf & Kaplan, 1972). Sets of 2–5 specific objectives were written for adjacent passage sentences. This permitted a single, generally phrased objective to be written for the same 2–5 adjacent sentences. Thus, a given set of 2–5 passage sentences was associated with either 2–5 specific objectives or with 1 general objective. The total number of specific objectives written for each of the three Length 56 passages was 36, 33, and 34 objectives. Combinations of these objectives were used for Lengths 113 and 169 such that the mean number of specific objectives per passage length was 34, 69, and 103, respectively. Similarly, the mean number of general objectives written for each of the Length 56 passages was 14, 12, and 8. The mean number of general objectives for Passage Lengths 56, 113, and 169 was 11, 23, and 34, respectively. The objectives always appeared in the same sequential order as the passage sentences.

Density was defined as the proportion of objective-relevant sentences to the total number of passage sentences. The use of all objectives (i.e., 103 specific or 34 general objectives) for Passage Length 169 resulted in a 60% density. That is, 103 objective-relevant sentences out of 169 passage sentences resulted in 60% of the passage sentences being objective-relevant. Subsets of these objectives were selected for Length 56 and 113 in order to maintain a 60% density across passage length. Table 1 shows the number of objectives and corresponding objective-relevant sentences used for Density 60%. A second density (40%) was achieved by selecting subsets of objectives from Density 60% such that 40% of the passage sentences were objective-relevant across all passage lengths (Table 1). Both density levels were constructed by selecting subsets of specific and matching general objectives that were relevant to the same passage sentences. Those sentences remaining, which were not relevant to any objective, were used to measure incidental learning. However, only those sentences which were identical in densities 40% and 60% (common sentences) were analyzed. This procedure permitted measurement of intentional and incidental learning with the same test items for every experimental condition.

All of the objectives for a given treatment were prepared in a list that preceded the entire passage. This whole presentation treatment permitted subjects to inspect all objectives and the passage concurrently. The objectives' lists were then divided into one third segments for part pres-

TABLE 1

Mean Number of Objectives and Passage Sentences per Treatment

Passage length (total passage sentences)	Density 40%			Density 60%		
	General objectives	Specific objectives	Objective-relevant passage sentences	General objectives	Specific objectives	Objective-relevant passage sentences
56	8	22	22	11	34	34
113	15	45	45	23	69	69
169	23	67	67	34	103	103

[3] Appreciation is extended to Mr. F. L. Stevenson for the use of materials from Systems Training Department courses at Bell Laboratories.

entation. Similarly, the passages were divided into one third segments corresponding to the objective segments. This procedure permitted inspection of the passage and corresponding objective segments. Both part and whole presentations were prepared for all experimental treatments.

Tests. The tests consisted of prompted recall, fill-in-the-blank type items. A test question was written for almost every passage sentence. Performance on objective-relevant test questions was considered a measure of intentional learning. Performance on nonobjective-relevant test questions was considered a measure of incidental learning. The total number of questions written for each Length 56 passage was 56, 52, 51. Combinations of these Length 56 test questions were used for Lengths 113 and 169. Three separate random orders of test questions were prepared for each passage length.

Analyses

Three 2 × 3 × 2 × 2 analyses of variance were performed with two levels of presentation (part and whole), three levels of passage length (56, 113, and 169 sentences), two levels of objective specificity (specific and general), and two levels of density (40% and 60%). Separate analyses were performed for intentional learning, incidental learning, and inspection time. Eighteen subjects were assigned to each of 24 treatments ($N = 432$). In addition, 108 subjects served in six reference groups who read the passages without objectives (18 subjects with part or whole texts per passage length).

Procedure

Approximately 100 subjects participated in each experimental session. The experimental sessions were conducted in each high school cafeteria after the last school period and lasted for about one and a half hours. The materials were packaged in manila envelopes with the contents of each envelope being inspected individually. The part treatment groups received five envelopes, each of which contained instructions and (*a*) the first one third of objectives and passage, (*b*) the second one third of objectives and passage, (*c*) the last one third of objectives and passage, (*d*) a test, and (*e*) supplementary material to occupy subjects who completed the experimental task early. The whole-treatment groups received three envelopes consisting of (*a*) a complete list of objectives and a complete passage, (*b*) a test, and (*c*) supplementary material. Half of the reference groups received the same three envelopes as the whole-treatment groups, except that there were no objectives in Envelope 1. The second half of the reference groups received the same five envelopes as the part-treatment groups, except that these envelopes contained no objectives. The experimental subjects were told that they would be tested only on the objective-relevant material. However, they were tested on almost every passage sentence. This procedure permitted measurement of both intentional (objective-relevant) and incidental (nonobjective-relevant) learning. All subjects were permitted to read at their own pace and were instructed to record their start and stop inspection times. Two digital clocks with 2 × 5 inch numerals were provided for this purpose.

Subjects. The subjects were 540 paid volunteers (181 males and 359 females between 15 and 18 years of age) from six New Jersey high schools.

Results

Intentional Learning

Figure 1 summarizes the means from several analyses. The data from the intentional learning analysis, for test items that were common to all treatments, is summarized in the top pair of curves in Figure 1. Arc sine transformations were used on proportion scores.

Part presentation ($\bar{X} = .39$) resulted in greater learning than whole presentation ($\bar{X} = .35$; $F = 4.49$, $df = 1/408$, $p < .05$). The passage length main effect was highly significant ($F = 23.04$, $df = 2/408$, $p < .001$). Paired comparisons using the New-

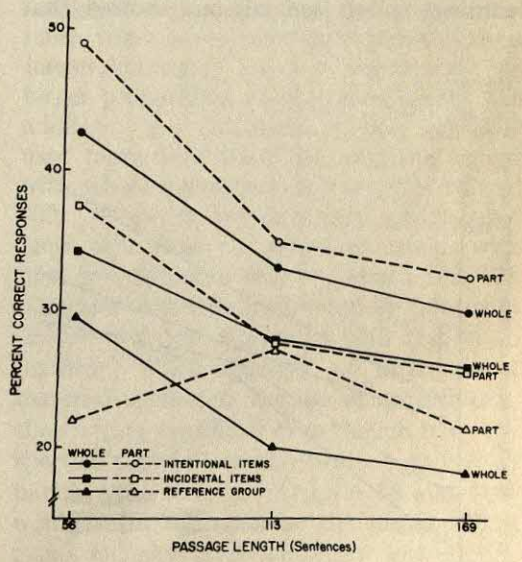


Figure 1. Mean proportion of correct responses for intentional and incidental test items, which were common to all treatments, as a function of passage length, and part versus whole presentations. (Nonobjective reference groups' proportion scores are also shown as a function of passage length and mode of presentation.)

man-Keuls technique showed that Length 56 ($\bar{X}$ = .46) resulted in greater learning than Length 113 ($\bar{X}$ = .34; q = 7.21; df = 408, r = 2, $p < .01$) and Length 169 ($\bar{X}$ = .31; q = 8.89, df = 408, r = 2, $p < .01$). Length 113 was not significantly different from Length 169. This finding held for both the part and whole presentations when analyzed separately. Neither specificity nor density had a significant effect.

Comparisons between the treatment groups and reference groups were made with t tests. The treatment groups' intentional learning was greater than the reference groups' learning for all three passage lengths, for both part and whole presentations.

Incidental Learning

The middle pair of curves in Figure 1 shows the main results for the analysis of incidental learning.

There was no significant difference for incidental learning between part and whole presentations. The passage length main effect was significant (F = 20.27, df = 2/408, $p < .001$). Similar to intentional learning, paired comparisons showed that performance with Length 56 ($\bar{X}$ = .36) was greater than with Length 113 ($\bar{X}$ = .28; q = 6.53, df = 408, r = 2, $p < .01$) and Length 169 ($\bar{X}$ = .26; q = 8.47, df = 408, r = 3, $p < .01$). Length 113 was not significantly different from Length 169. Neither specificity nor density had a significant effect.

The treatment groups' incidental learning was significantly greater than the reference groups' incidental learning with part presentations for Length 56 only (t = 3.79, df = 88, $p < .01$). No significant difference was found for Lengths 113 or 169. However, the treatment groups' incidental learning was significantly greater than the reference groups' incidental learning with whole presentations for Lengths 113 and 169 (t = 2.87, df = 88, $p < .01$, and t = 3.00, df = 88, $p < .01$, respectively). No significant difference was found for Length 56.

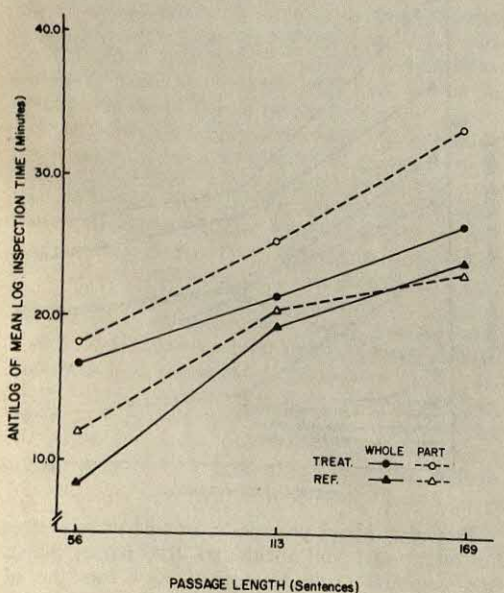


FIGURE 2. Anti-log of mean log inspection time for part versus whole presentations for (TREAT.) treatment and REF. (reference) groups as a function of passage length.

Inspection Time

Some of the time data were unusable due to subject errors in recording times. Therefore, data for only 450 subjects were analyzed (treatment groups: n = 360; reference groups: n = 90). Log transformations were performed on inspection time. Figure 2 shows the main results of this analysis.

Passages and objectives with part presentations ($\bar{X}$ = 25.76) required more reading time than with whole presentations ($\bar{X}$ = 21.58; F = 6.71, df = 1/336, $p < .01$). Shorter passages required less reading time than longer passages (Length 56, $\bar{X}$ = 17.48; Length 113, $\bar{X}$ = 23.54; Length 169, $\bar{X}$ = 29.99; F = 4.10, df = 2/336, $p < .001$). General objectives ($\bar{X}$ = 21.55) required less time than specific objectives ($\bar{X}$ = 25.78; F = 13.85, df = 1/336, $p < .001$). Density 40% ($\bar{X}$ = 21.93) required less time than Density 60% ($\bar{X}$ = 25.45; F = 8.47, df = 1/336, $p < .01$). None of the interactions was significant.

Inspection time comparisons between each reference group and the corresponding treatment group were made with t tests. The only significant difference in the part presentation occurred at Length 169 where the treatment group ($\bar{X}$ = 33.21) required more time than the reference group ($\bar{X}$ = 22.65; t = 2.31, df = 33, $p < .05$). The

treatment group ($\bar{X} = 16.63$) also required more time than the reference group ($\bar{X} = 8.07$) with the whole presentation at Length 56 ($t = 2.59$, $df = 73$, $p < .01$). In addition, the treatment group ($\bar{X} = 26.77$) required more time than the reference group ($\bar{X} = 24.11$) with the whole presentation at Length 169 ($t = 2.76$, $df = 73$, $p < .01$).

Discussion

The primary finding of this study was that performance increments found with whole presentations of instructional objectives could be achieved with part presentations of objectives and text. In fact, the part presentations resulted in even greater intentional learning than whole presentations.

As expected, intentional learning with whole presentations was greater than the reference groups' performance on the same test items for every passage length. This finding is consistent with the results of the Kaplan and Rothkopf (1974) study. Similarly, the greater intentional learning found with part presentations over the reference groups' performance at every passage length is consistent with the findings of Frase (1967) and Rothkopf and Bisbicos (1967). These studies found intentional learning to be greater than a nonobjective control group when adjunct questions were distributed throughout a text. Incidental learning with whole presentations was found to be greater than the reference groups' performance at Passage Lengths 113 and 169 but not at Passage Length 56. This finding also replicates Kaplan and Rothkopf (1974). Conversely, incidental learning with part presentations was found to be greater than the reference groups' performance only at Length 56. This was the result of a reversal in the reference groups' performance with part and whole presentations (Figure 1). That is, the reference groups' performance was greater with whole presentations at Length 56 and with part presentations at Lengths 113 and 169. This was predictable in that learning has been shown to be greater with whole presentations of short contextually connected passages and with part presentations of longer, less connected passages (Ausubel, 1963). Conversely, no part/whole effect was found for the treatment groups' incidental learning. Both the treatment groups' performance on incidental items and the reference groups' total performance are based upon nonobjective-relevant text sentences. However, the reference groups' performance is considered to be intentional learning in that they were instructed to learn everything in the passage. Thus, the reference groups read the passage nonselectively, giving equal attention to each sentence. The treatment groups, on the other hand, read the text selectively giving less attention to incidental sentences.

Both intentional and incidental learning for part and whole presentations decreased as a function of passage length. This finding, with respect to whole presentation, is again consistent with the Kaplan and Rothkopf (1974) study. Similar to that study and to Rothkopf & Kaplan (1972) was the finding that differences in specificity of stating objectives and in density were in the same direction (i.e., specific > general objectives and Density 40% > Density 60%), although the differences were not significant.

The inspection time analysis shows that more time is used for part presentations, longer passages, specific objectives, and larger proportions of objectives to text. In addition, the treatment groups generally used more time than the reference groups with whole presentations (Lengths 56 and 169). However, the treatment groups used more time than the reference groups with part presentations only at Length 169. The additional time generally used by treatment groups must be considered with respect to learning. When mastery of instructional material is desired, the use of objectives is shown to be beneficial even though more inspection time is needed. Smaller segments of partial presentations (Lengths 56 and 113) would seem to maximize the likelihood of mastering any given objective while minimizing inspection time differences.

REFERENCES

Ausubel, D. P. *The psychology of meaningful verbal learning*. New York: Grune & Stratton, 1963.

Frase, L. T. Learning from prose material: Length

of passage, knowledge of results and position of question. *Journal of Educational Psychology*, 1967, **58**, 266–272.

Kaplan, R., & Rothkopf, E. Z. Instructional objectives as directions to learners: Effects of passage length and amount of objective-relevant content. *Journal of Educational Psychology*, 1974, **66**, 448–456.

Papay, J. P. An investigation of the effects of type, location, and distribution of learning instructions on the acquisition and retention of meaningful prose materials. Unpublished doctoral dissertation, Florida State University, 1971.

Rothkopf, E. Z., & Bisbicos, E. E. Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology*, 1967, **58**, 36–61.

Rothkopf, E. Z., & Kaplan, R. Exploration of the effect of density and specificity of instructional objectives on learning from text. *Journal of Educational Psychology*, 1972, **63**, 295–302.

(Received December 6, 1973)

(*Continued from p. 709*)

Effects of Vicarious Consequences and Model and Experimenter Sex on Imitative Behavior in First-grade Children. Candace S. Garrett (Department of Educational Psychology, Indiana University, University School 142, 2805 E. Tenth Street, Bloomington, Indiana 47401) and Donald J. Cunningham.

Relationship between Mental Abilities, Social Class, and Exposure to English in Chinese Fourth Graders. Leland Y. Lee (Department of Psychology, University of Hawaii, 2430 Campus Road, Honolulu, Hawaii 96822) and Rolfe LaForge.

Learning Ability and Intelligence as Cognitive Predictors of Achievement in First-grade Black Children. Trevor E. Sewell (Department of School Psychology, Temple University, Philadelphia, Pennsylvania 19122) and Roger A. Severson.

An Exploratory of Two Correction Procedures Used in Mastery Learning Approaches to Instruction. James H. Block (Department of Education, University of California, Santa Barbara 93106) and Michael L. Tierney.

Facilitated Learning in Connected Discourse. James M. Royer (Department of Psychology, University of Massachusetts, Amherst, Massachusetts 01002) and Glen W. Cable.

Children' Solution Processes in Arithmetic Word Problems. Daniel A. J. Rosenthal (Department of Educational Psychology, College of Education, Temple University, Philadelphia, Pennsylvania 19122) and Lauren B. Resnick.

Effects of Mode of Modeling and Model Age and Ethnicity on Rule-governed Language Behaviors. Mary B. Harris (Department of Educational Foundations, College of Education, University of New Mexico, Albuquerque, New Mexico 87106) and James I. Grieshop.

Use of Rational Concepts by Inner-city Children. Judith A. Meissner (Developmental Research Division, Educational Testing Service, Princeton, New Jersey 08540).

Classroom Seating and Functional Brain Asymmetry. Ruben C. Gur (Department of Psychology, Stanford University, Stanford, California 94305), Raquel E. Gur, and Brachia Marshalek.

A Mnemonic Method for Learning a Second Language Vocabulary. Richard C. Atkinson (Department of Psychology, Stanford University, Stanford, California 94305) and Michael R. Raugh.

Further Comparison of Imagery and Vocalization Strategies in Children's Discrimination Learning. Joel R. Levin (University of Wisconsin, 1025 West Johnson Street, Madison, Wisconsin 53706), Elizabeth S. Ghatala, Thomas M. DeRose, Larry Wilder, and Richard W. Norton.

Fixed and Ascending Criteria for Unit Mastery Learning. John G. Carlson (Department of Psychology, University of Hawaii, 2430 Campus Road, Honolulu, Hawaii 96822) and Karl A. Minke.

Syllogistic Reasoning: The Generation of Counterexamples. Fred Helsabeck, Jr. (Faculty of Natural Sciences and Mathematics, Stockton State College, Pomona, New Jersey 08240).

Influence of Sampling and Comparison Processes on the Development of Communication Effectiveness. Steven R. Asher (Department of Educational Psychology, 188 Education Building, University of Illinois, Urbana, Illinois 61801) and Ross D. Parke.

Journal of Educational Psychology
1974, Vol. 66, No. 5, 793–799

COGNITIVE STYLE AS A VARIABLE IN SCHOOL LEARNING

JACK E. ROBINSON[1] AND JERRY L. GRAY

University of Iowa

The purpose of the study was to investigate the relationship between cognitive style and school learning of fifth-grade children. The findings indicate that cognitive style was differentially related to school learning for both boys and girls after variance attributed to verbal and nonverbal IQ was taken into consideration in the data analysis. The results suggest that the relationship between a particular style and a particular school-learning variable may be an important consideration prior to assigning children to differential instructional treatments or to instructing children in the use of a particular style.

The possible effects of individual differences in cognitive style on school learning has been speculated upon since the notion of cognitive style was introduced by Gardner 20 years ago (Gardner, 1953; Glaser, 1972; Kagan, Moss, & Sigel, 1963; Messick, 1970; Nunney & Hill, 1972; Wallach & Kogan, 1965; Witkin, Dyk, Faterson, Goodenough, & Karp, 1962). Empirical evidence, however, has led to few conclusions concerning the relationships among cognitive style, learning conditions, and learning outcomes (Coop & Sigel, 1971; Cronbach, 1968; Kogan, 1971).

Although the relationship between cognitive style and school learning may be masked because individual differences in style are not commonly matched with learning conditions, a more basic problem may lie in style tests themselves. Among the problems most often cited are scoring based on ipsative formats and the irrelevance of performance on style tests in relation to school learning tasks (Annesley, 1971; Brozovich, Hall, & Watson, 1972; Davis, 1971; Denmark, Havlena, & Murgatroyd, 1971; Denney, 1971; Gatewood, 1971; Huckabee, 1969; Scott, 1971; Wallach & Kogan, 1965).

The purpose of the present study was to investigate further the relationship between cognitive style and school learning of fifth-grade children. A test was designed for this study which featured independent scales for each cognitive style and tasks similar to school-learning tasks. Performance on a standardized achievement test was used as a measure of school learning. Since past studies have suggested that interactions among cognitive style, intellectual ability, and sex may mask the relationship between cognitive style and school learning, intellectual ability and sex were included as additional variables for investigation. Two general questions were used to guide the data analysis. First, what is the relationship between cognitive style and school learning? Second, what additional contribution do measures of cognitive style add beyond that obtained from a traditional test of intelligence in the prediction of school learning?

The cognitive style dimension under study was identified by Kagan et al. (1963). Kagan et al. defined cognitive style as "stable individual preferences in modes of perceptual organization and conceptual categorization of the external environment [p. 74]." These investigators subsequently identified three cognitive styles among children: categorical, descriptive, and relational. Categorical responses refer to the use of common class membership in relating stimuli (e.g., a dog and a sheep are both animals). This style has been referred to as an inferential mode of conceptualization in that the use of abstract labels is a means

[1] Requests for reprints should be sent to Jack E. Robinson, who is now at the Department of Educational Foundations, Old Dominion University, Norfolk, Virginia 23508.

of summarizing the detailed relationship among stimuli. Descriptive responses may be defined as concepts formed on the basis of shared physical attributes of stimuli (e.g., a dog and a sheep both have four legs). This style has been called an analytical mode of conceptualization in the sense that an individual deals with similarities among the concrete detail of stimuli. Relational responses are those in which functional relationships among stimuli are used in associating stimuli (e.g., a dog is used to drive sheep). The relational style has been referred to as a global, contextual, thematical mode in that an individual associates a whole stimulus with another whole stimulus in making an interdependent functional relationship rather than forming a concept. The definitions of each of these styles, characteristics of items on tests, and scoring criteria for classifying items from earlier investigations were used as the basis for designing the instrument of the present study (Achenbach, 1970; Brozovich et al., 1972; Kagan et al., 1963; Wallach and Kogan, 1965).

Method

Subjects

Two-hundred and fifty-eight fifth grade children (132 boys, 126 girls) from 12 classrooms in five schools in a midwestern city served as subjects.

Tests

The Iowa Test of Basic Skills (ITBS) and the Lorge-Thorndike Intelligence Test were administered as a regular part of the school district's evaluation program. Verbal and nonverbal deviation IQ scores for the Lorge-Thorndike Intelligence Test and grade-equivalent scores for each subtest of the ITBS were subsequently obtained by the experimenters.

The cognitive style test consisted of verbal analogy items.[2] Each cognitive style scale (categorical, descriptive, and relational) included 14 items. The cognitive style test was administered orally by the subjects' classroom teachers and scored by the experimenters.

Verbal content was chosen for the cognitive style test since most school-related tasks are predominantly verbal. An analogy format was chosen for the test as this type of format provides a means of assessing cognitive style as a process variable. That is, a subject must form a relationship between the first two terms and then generalize the relationship between the first two terms to the last two terms of the analogy. A separate scale was constructed for each style in order to derive independent scores for each subject. Most instruments used in this research area have consisted of pictorial content, items involving a two-term association, and ipsative or intraindividual dependent scores for subjects.

A basic assumption underlying most research has been that cognitive style is an individual's manner rather than one's level of intellectual functioning (Kogan, 1971). Consequently, precautions were taken to select words for expressing analogous relationships which were familiar to fifth-grade subjects.

Reliabilities (Kuder-Richardson Formula 20) for the scales were categorical (C) .84; descriptive (D) .80; and relational (R) .78 for the 258 subjects. Thus, subjects performed fairly consistently on each of the 14 item scales.

Results and Discussion

Means and standard deviations of style, IQ, and school-learning variables are reported in Table 1 for boys, girls, and the total sample. Cognitive style scores are given in raw score form, verbal and nonverbal IQ scores in deviation IQ units, and school learning scores in grade-equivalent

---

[2] The following are examples of each type of item (* indicates the correct alternative):

Categorical:
Dog is to cat as chicken is to _____.
a. feather b. eggs *c. pig d. bark

Descriptive:
Chair is to legs as lamp is to _____.
a. furniture b. light c. hand
*d. light bulb

Relational:
Key is to lock as saw is to _____.
a. keys *b. board c. tool d. teeth

TABLE 1
DESCRIPTIVE STATISTICS

Variable	Subjects					
	Total (n = 258)		Boys (n = 132)		Girls (n = 126)	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Categorical style	9.8	3.6	9.5	3.7	10.2	3.5
Descriptive style	9.2	3.4	8.7	3.4	9.8	3.3
Relational style	9.3	3.2	9.0	3.3	9.7	3.1
Verbal IQ	99.4	14.1	97.0	14.2	102.0	13.6
Nonverbal IQ	107.3	16.7	105.8	16.6	108.9	16.6
Vocabulary	4.5	1.5	4.2	1.6	4.9	1.5
Reading Comprehension	4.4	1.4	4.1	1.3	4.8	1.4
Spelling	4.4	1.6	4.0	1.4	4.9	1.7
Capitalization	4.7	1.7	4.4	1.4	5.0	1.8
Punctuation	4.5	1.7	4.0	1.4	5.0	1.8
Language Usage	4.3	1.6	4.1	1.5	4.6	1.7
Map Reading	4.5	1.3	4.5	1.3	4.6	1.3
Graphs and Tables	4.6	1.4	4.5	1.4	4.6	1.5
Reference Materials	4.5	1.4	4.2	1.2	4.8	1.5
Mathematical Concepts	4.6	1.2	4.5	1.1	4.6	1.2
Mathematical Problems	4.5	1.2	4.4	1.2	4.7	1.3

units. An inspection of the means and standard deviations indicates that they were fairly uniform across sex for all variables. While girls obtained slightly higher means on all variables, the standard deviations were approximately the same for each sex.

Table 2 presents the correlation coefficients among style, IQ, and school-learning variables for the total sample. When the same set of correlations were calculated separately for each sex, there were only two significantly different correlation coefficients. The correlations between reading comprehension and language usage and between descriptive style and vocabulary were significantly higher ($p < .05$) for girls than boys.

A series of multiple regression analyses was performed to study the unique contribution of each cognitive style measure in the prediction of school learning. The results of this series of analyses are reported in Table 3 for each sex. Six of the ITBS subtests (Vocabulary, Reading Comprehension, Capitalization, Map Reading, Graphs and Tables, and Mathematical Concepts) involved multiple cognitive style orientations based on significant style contributions across sex. The five remaining subtests were less complex in cognitive style requirements.

Descriptive style contributed significantly in 19 of the 22 analyses for both sexes. A reversal occurred, however, for categorical and relational styles. For boys, relational style accounted for 7 of 10 additional significant style contributions. Seven of the 9 significant style contributions for girls were for categorical style.

The only school-learning variable for which a style other than descriptive entered the regression equation first for both sexes was mathematical concepts. Categorical style was the first style to enter the equation for this subtest. In the other two instances (Reading Comprehension and Graphs and Tables) in which descriptive style did not enter the regression equation first, relational style entered first for boys. These were the only two subtests in which relational style made a significant contribution in relation to school learning for girls.

An additional series of multiple regression analyses was performed in order to investigate the unique contribution of cognitive style in the prediction of school learning after the effects of verbal and nonverbal IQ were removed from the school-learning variables. Fifteen independent variables, including 10 interaction terms (e.g., Descriptive × Verbal IQ), were used in this series of

TABLE 2

INTERCORRELATIONS AMONG ALL VARIABLES

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Categorical style															
2. Descriptive style	.68														
3. Relational Style	.78	.75													
4. Verbal IQ	.58	.59	.56												
5. Nonverbal IQ	.62	.60	.59	.72											
6. Vocabulary	.51	.54	.51	.69	.51										
7. Reading Comprehension	.50	.54	.54	.70	.57	.81									
8. Spelling	.52	.57	.47	.71	.56	.69	.70								
9. Capitalization	.47	.52	.44	.63	.59	.58	.64	.68							
10. Punctuation	.48	.55	.46	.63	.55	.60	.69	.79	.78						
11. Usage	.45	.56	.45	.58	.51	.69	.71	.67	.67	.69					
12. Map Reading	.52	.53	.51	.64	.68	.61	.67	.60	.60	.62	.59				
13. Graphs and Tables	.46	.46	.47	.56	.58	.61	.67	.51	.56	.55	.54	.69			
14. Reference Materials	.46	.51	.45	.65	.59	.68	.76	.67	.66	.72	.63	.70	.68		
15. Mathematical Concepts	.56	.54	.51	.64	.69	.56	.65	.62	.61	.63	.57	.66	.61	.63	
16. Mathematical Problems	.51	.55	.48	.68	.65	.55	.63	.67	.66	.68	.57	.63	.62	.67	.74

Note. All correlation coefficients are significant at the .001 level.

TABLE 3

RESULTS OF MULTIPLE REGRESSION ANALYSES OF COGNITIVE STYLES IN RELATION TO SCHOOL LEARNING

Variable	Sex	Proportion of Variance					
		Removed by most potent style variable			Attributed to second style variable entered[a]		
		Style	r	r^2	Style	Increase in R^2	F
Vocabulary	boy	D	.44	.19	R	.03	4.92*
	girl	D	.63	.40	C	.04	8.04**
Reading Comprehension	boy	R	.51	.26	C	.03	5.60*
	girl	D	.58	.34	R	.02	4.28*
Spelling	boy	D	.52	.27			
	girl	D	.59	.38	C	.03	6.27*
Capitalization	boy	D	.56	.31	R	.02	4.52*
	girl	D	.48	.23	C	.03	4.61*
Punctuation	boy	D	.50	.25	R	.02	4.31*
	girl	D	.57	.33			
Language Usage	boy	D	.51	.26			
	girl	D	.59	.34			
Map Reading	boy	D	.58	.34	R	.05	9.34**
	girl	D	.48	.23	C	.03	5.57*
Graphs and Tables	boy	R	.50	.25	C	.04	7.92**
	girl	D	.45	.20	R	.03	4.16*
Reference Materials	boy	D	.40	.16			
	girl	D	.58	.34			
Mathematical Concepts	boy	C	.54	.29	R	.05	10.00**
	girl	C	.58	.34	D	.03	5.70*
Mathematical Problems	boy	D	.55	.30			
	girl	D	.54	.29	C	.03	6.24*

Note. Abbreviations: D = descriptive, R = relational, and C = categorical.

[a] In no instance did the entry of a third style measure contribute significantly in the prediction of school learning at the .05 level.

* Significant at .05 level.

** Significant at .01 level.

analyses. For each analysis, verbal and nonverbal IQ were forced into the equation, then the remaining 13 independent variables entered freely into the regression equation. The results of this series of analyses are presented in Table 4. In order to interpret significant interactions between style and IQ, the zero-order correlations between cognitive style and school learning for different IQ levels were calculated. The interactions between IQ and a particular style were relatively consistent across sex and school-learning variables. The relationships between relational style and reading comprehension, map reading, and graphs and tables were higher in the upper-third IQ level than the lower two-thirds IQ levels for boys. For categorical style, the relationship between style and mathematical concepts was higher in the upper two-thirds IQ levels than the lower-third IQ level for both sexes. The relationships between descriptive style and the other school-learning measures were higher in the upper-third and lower-third IQ levels than in the middle-IQ level for both sexes. The only exceptions to these trends were that the relationships between descriptive style and punctuation and descriptive style and language usage were higher for middle-IQ level than high- and low-IQ levels for girls. A comparison of Tables 3 and 4 indicates that the relationships between relational style and school learning for boys and between descriptive style and school learning for both sexes were still present after the effects of verbal and nonverbal IQ were partialed out of the school-learning measures. The relationship

TABLE 4

CONTRIBUTION OF COGNITIVE STYLES AFTER VARIANCE ATTRIBUTED TO VERBAL (V) AND NONVERBAL (NV) IQ HAS BEEN REMOVED FROM SCHOOL LEARNING

Variable	Sex	Proportion of variance				
		Removed by V IQ and NV IQ		Attributed to additional variables		
		R	R^2	Variable entered	Increase in R^2	F
Vocabulary	boy	.66	.43			
	girl	.71	.50	D × V IQ	.07	21.41**
Reading Comprehension	boy	.65	.43	R × V IQ	.05	11.00**
				R	.04	9.21**
	girl	.75	.56	D × V IQ	.03	9.30**
				D	.02	4.90*
Spelling	boy	.68	.47	D × NV IQ	.02	4.88*
	girl	.73	.54	D × NV IQ	.04	11.42**
Capitalization	boy	.67	.45	D × NV IQ	.04	9.78**
	girl	.65	.42			
Punctuation	boy	.61	.38	D × NV IQ	.02	5.26*
	girl	.66	.44	D × NV IQ	.05	12.53**
Language Usage	boy	.56	.32	D × V IQ	.06	11.48**
				D	.03	5.74*
	girl	.61	.37	D × NV IQ	.09	19.35**
				D	.02	5.02*
Map Reading	boy	.72	.52	R × V IQ	.03	10.06**
				R	.02	5.40*
	girl	.71	.50			
Graphs and Tables	boy	.64	.41	R × V IQ	.03	7.07**
				R	.04	9.70**
	girl	.60	.36			
Reference Materials	boy	.69	.47			
	girl	.65	.42	D × NV IQ	.06	14.74**
				D	.02	4.59*
Mathematical Concepts	boy	.74	.54	C × NV IQ	.02	6.34*
				C	.03	8.66**
	girl	.72	.52	C × NV IQ	.04	9.69**
				C	.04	12.40**
Mathematical Problems	boy	.70	.50	D × NV IQ	.03	7.28**
	girl	.74	.55			

Note. Abbreviations: D = descriptive, R = relational, and C = categorical.
* Significant at .05 level.
** Significant at .01 level.

between categorical style and school learning for girls was reduced considerably after the variance attributed to verbal and nonverbal IQ were removed from school learning.

SUMMARY

The results of this study support earlier contentions that standardized intellectual ability and school achievement tests are quite heterogeneous with respect to cognitive style requirements. The findings indicate, however, that additional variance was accounted for by cognitive style measures beyond that of verbal and nonverbal IQ in relation to school learning.

Of the three style measures, descriptive style was the most important contributor in the prediction of school learning. For boys, relational style contributed more often in the prediction of school learning than a categorical style. A categorical style accounted for additional variance in the prediction of school learning more frequently for girls than a relational style. The contribution of categorical style for girls, how-

ever, was reduced substantially after variance attributed to verbal and nonverbal IQ was removed from school-learning variables.

The high intercorrelations among the three style scales for the instrument used in this study would suggest difficulty in obtaining differential convergent validity for cognitive style. Despite these high intercorrelations, the differential relationships between cognitive style and school learning found in this investigation indicate that convergent validity can be demonstrated for the notion of cognitive style in relation to school learning. Whether or not reducing the intercorrelations among scales would allow for greater clarification of the relationship between cognitive style and school learning than found in this study is a question which warrants further investigation.

Based on the findings of this study, it would seem that before meaningful recommendations can be made for educational programming based on the cognitive style mapping of children, the relationship between a particular cognitive style and a particular school-learning task must be taken into consideration.

REFERENCES

Achenbach, T. M. The children's associative responding test: A possible alternative to group IQ tests. *Journal of Educational Psychology,* 1970, **61,** 349–357.

Annesley, F. R. Cognitive style as a variable in the reading achievement and intelligence of boys. *Dissertation Abstracts International,* 1971, **32,** 1910-A.

Brozovich, R., Hall, L. P., & Watson, R. Development of a scale to evaluate cognitive styles and abilities. (Washington, D.C.: U.S. Office of Education, Project No 1-E-037), 1972.

Coop, R. H., & Sigel, I. E. Cognitive style: Implications for learning and instruction. *Psychology in the Schools,* 1971, **8,** 152–161.

Cronbach, L. J. Intelligence? Creativity? A parsimonious reinterpretation of the Wallach-Kogan data. *American Educational Research Journal,* 1968, **5,** 491–511.

Davis, A. J. Cognitive style: Methodological and developmental considerations. *Child Development,* 1971, **42,** 1447–1459.

Denmark, F. L., Havlena, R. A., & Murgatroyd, D. Reevaluation of some measures of cognitive styles. *Perceptual and Motor Skills,* 1971, **33,** 133–134.

Denney, D. The assessment of differences in conceptual style. *Child Study Journal,* 1971, **1,** 142–155.

Gardner, R. W. Cognitive styles in categorizing behavior. *Journal of Personality,* 1953, **22,** 214–233.

Gatewood, M. M. Cognitive styles and academic performance. *Dissertation Abstracts International,* 1971, **29,** 4417-A.

Glaser, R. Individuals and learning: The new aptitudes. *Educational Researcher,* 1972, **1,** 5–12.

Huckabee, M. W. Personality and academic aptitude correlates of cognitive control principles. *Southern Journal of Educational Research,* 1969, **3,** 1–9.

Kagan, J., Moss, H. A., & Sigel, I. E. Psychological significance of styles of conceptualization. *Society for Research in Child Development Monographs,* 1963, **28,** 73–112.

Kogan, N. Educational implications of cognitive styles. In G. S. Lesser (Ed.), *Psychology and educational practice.* Glenview, Ill.: Scott, Foresman, 1971.

Messick, S. The criterion problem in the evaluation of instruction: Assessing possible not just intended outcomes. In M. C. Wittrock & D. E. Wiley (Eds.), *The evaluation of instruction: Issues and problems.* New York: Holt, Rinehart & Winston, 1970.

Nunney, D. K., & Hill, J. E. Personalized educational programs. *Audiovisual Instruction,* 1972, **17** (2), 10–15.

Scott, N. C. Cognitive style assessment: One test or several. Paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

Wallach, M. A., & Kogan, N. *Modes of thinking in young children.* New York: Holt, Rinehart & Winston, 1965.

Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R., & Karp, S. A. *Psychological Differentiation.* New York: Wiley, 1962.

(Received December 26, 1973)

In addition to the members of the Advisory Board, the *Journal of Educational Psychology* asks other psychologists to aid in the review of manuscripts. In this way we have been able to achieve a broader review base. The Editors wish to thank the following people for their generous assistance to the Journal.

WILLIAM H. ANGOFF, *Educational Testing Service*
RICHARD ANDERSON, *University of Illinois*
DONALD BAER, *University of Kansas*
LEONARD L. BAIRD, *Educational Testing Service*
URSALA BELLUGI, *Salk Institute*
LEONARD BLACKMAN, *Teachers College, Columbia University*
MARION BLANK, *Yeshiva University*
* BENJAMIN BLOOM, *University of Chicago*
GORDON BOWER, *Stanford University*
EDWARD CALDWELL, *West Virginia University*
ROBERT CALFEE, *Stanford University*
JOHN B. CARROLL, *Educational Testing Service*
RONALD CARVER, *American Institutes for Research*
COURTNEY CAZDEN, *Harvard University*
WILLIAM CRANO, *University of Michigan*
JOHN E. CENTRA, *Educational Testing Service*
WILLIAM W. COOLEY, *University of Pittsburgh*
* JAMES CROUSE, *University of Texas*
ROBYN DAWES, *University of Oregon*
ORPHA KAY DUELL, *Kansas State University*
SHEL FELDMAN, *Swarthmore College*
KAREN FOX, *Teachers College, Columbia University*
NORMA FURST, *formerly of Temple University*
JACOB L. GERWITZ, *National Institute of Mental Health*
LILA GLEITMAN, *University of Pennsylvania*
JAMES GREENO, *University of Michigan*
* JOHN GUTHRIE, *John F. Kennedy Institute*
GORDON A. HALE, *Educational Testing Service*
RODNEY T. HARTNETT, *Educational Testing Service*
BETTY J. HOUSE, *University of Connecticut*
DONALD HUBBARD, *University of Minnesota*
MICHAEL S. HUMPHREYS, *Northwestern University*
RICHARD M. JAEGER, *University of South Florida*
MARTIN KAUFMAN, *Department of Education, U.S. Department of Health, Education, and Welfare*
THOMAS LANDAUER, *Bell Laboratories*
* ALAN LESGOLD, *University of Pittsburgh*
RICHARD MAYER, *University of Indiana*
DONALD MACMILLAN, *University of California, Riverside*
DONALD MEICHENBAUM, *University of Waterloo*
LEONARD MILLER, *University of Iowa*
RICHARD T. MURPHY, *Educational Testing Service*
EDITH NEIMARK, *Rutgers University*
IRENE NICHOLS, *Northwestern University*
CALVIN NODINE, *Temple University*
JOHN NOLAN, *Teachers College, Columbia University*
ALAN PAIVIO, *University of Western Ontario*
DONALD A. ROCK, *Educational Testing Service*
* ROBERT R. ROSENTHAL, *Harvard University*
KEITH RAYNOR, *University of Rochester*
JOSEPH SCANDURA, *University of Pennsylvania*
THOMAS SHUELL, *State University of New York at Buffalo*
R. J. SHAVELSON, *Stanford University*
DAVID SEARS, *University of California at Los Angeles*
HAROLD W. STEVENSON, *University of Michigan*
BETH SULZER-AZAROFF, *University of Conneticut*
THOMAS STICHT, *H. R. O.*
RICHARD TUCKER, *McGill University*
RICHARD VENEZKY, *University of Wisconsin*
RICHARD WAITE, *Denver, Colorado*
JAMES YOUNISS, *Catholic University of America*

* Reviewed more than one manuscript.

A New Book! *Reinert*

CHILDREN IN CONFLICT: Educational Strategies

An overview of the field of the behaviorally disturbed child in education, this text presents a variety of theoretical approaches to the subject. Psychological, behavioral, biogenetic, sociological and ecological aspects are considered. Although the emphasis of the text is on the mainstream of education, the regular classroom, other delivery systems are discussed when appropriate. Thorough discussions are included on: screening and diagnostic procedures, placement of children, and programming for the individual child. A glossary of terms, information on films and other teaching materials is also included.

By Henry R. Reinert, Ed.D. May, 1975. Approx. 275 pages, 7″ × 10″, 20 illustrations. About $8.95.

Gearheart

LEARNING DISABILITIES: Educational Strategies

Help your students face the perplexing problems of the learning disabled child. This comprehensive overview of the learning disabilities field prepares them for the task with discussions on: perceptual motor theories, screening for learning disabilities, behavior modification, medical intervention, and strategies for developing successful learning disabilities programs. Actual case studies illustrate various learning disabilities.

By B. R. GEARHEART. Foreword by ERNEST P. WILLENBERG. 1973, 234 pages plus FM I–XIV, 7″ × 10″, illustrated, Price, $9.50.

INSTRUCTORS NOTE: For a complimentary copy to evaluate firsthand for possible class adoption, simply write to the Textbook Department. Please be specific as to your course and the applicability of the requested text therein. (Offer good only in U.S. and Canada.)

MOSBY

TIMES MIRROR

THE C. V. MOSBY COMPANY
11830 WESTLINE INDUSTRIAL DRIVE
ST. LOUIS, MISSOURI 63141



The 1st Annual

TEMPLE CONFERENCE

on

BEHAVIOR THERAPY AND BEHAVIOR MODIFICATION

50 Years of Progress

November 14–17, 1974

Lectures and workshops at introductory and advanced levels.
Partial list of lecturers and workshop leaders:

Teodoro Ayllon	Paula Bram	Iris Goldstein Fodor	Debora Phillips
L. Michael Ascher	Joseph R. Cautela	Mary Cover Jones	Leo J. Reyna
Nathan Azrin	Edna Foa	Ogden Lindsley	Thomas Stampfl
Donald Baer	Martin Gittelman	Richard W. Malott	Johann Stoyva
David Barlow	Israel Goldiamond	Neil E. Miller	Joseph Wolpe
John P. Brady	Alan Goldstein	Peter Nathan	

Sponsored by the Behavior Therapy Unit, Department of Psychiatry, Temple University School of Medicine.

FEES: Pre-registration—$35.00 ($40.00 after Oct. 20, 1974) for the general sessions. There will be an additional fee for workshops and conversation hours. The number of participants for each workshop and conversation hour will be limited.

For additional information please write or call:

Dr. Edna Foa
Behavior Therapy Unit
Temple Department of Psychiatry, Room 709
c/o E.P.P.I.
Henry Avenue at Abbottsford Road
Philadelphia, Pa. 19129
215-GE-8-4298

39 years of continuous service

for Discriminating Test Users

Edited by OSCAR K. BUROS

NEW Tests in Print II.

In press — to be published late 1974 — a comprehensive index to tests, test reviews, and the literature on specific tests; approximately 1,300 two-column pages. TIP II includes a bibliography of approximately 2500 tests published in English as of March 1974; over 16,000 new references through 1971 on the construction, use, and validity of specific tests; a cumulative name index for each test with references; a complete title index of all published tests (including out of print); and a name index of all authors of tests, reviews, excerpts, and references in the *Mental Measurements Yearbook* series.

1972 The Seventh Mental Measurements Yearbook

1972; 2032 pages in 2 volumes; $70 per set. "Indispensable . . . comprises all the values of its predecessors, but in larger measure . . . simple, clear indexing system . . . statistics can give only the most inadequate picture . . . the newest edition lists 1,157 tests with 12,372 references. For 546 of the tests, there are 798 original reviews by 439 reviewers, supplemented by another 181 reviews excerpted from 39 journals." — *Test Service Bulletin of The Psychological Corporation,* May 1972.

1970 Personality Tests and Reviews

1970; 1690 pages; $45.00. "The second of four monographs covering specific sections of the first six *Mental Measurements Yearbooks.* * A great deal of new material is also presented for the first time." — Wayne H. Holtzman, *Am J Orthopsychiatry,* July 1970.

1968 Reading Tests and Reviews

1968; 542 pages; $20.00. "Brings together in convenient form the enormous amount of bibliographical material about reading tests of the past and present that have been available heretofore only by consulting separately the six *Mental Measurements Yearbooks.*" — Frederick B. Davis, *Ed & Psychol Meas,* Autumn 1969.

Earlier Volumes Needed for Full Coverage

1965 **The Sixth Mental Measurements Yearbook.** Reprinted 1971; 1751 pages; $45.00.
1961 **Tests in Print.** Reprinted 1974; 508 pages; $15.00.
1959 **The Fifth Mental Measurements Yearbook.** 1321 pages; $35.00.
1953 **The Fourth Mental Measurements Yearbook.** Reprinted 1974; 1188 pages; $30.00.
1949 **The Third Mental Measurements Yearbook.** Reprinted 1974; 1062 pages; $25.00.
1941 **The 1940 Mental Measurements Yearbook.** Reissued 1972; 672 pages; $22.50.
1938 **The 1938 Mental Measurements Yearbook.** Reissued 1972; 430 pages; $20.00.

Order directly from
THE GRYPHON PRESS 220 Montgomery Street, Highland Park, N. J. 08904

Thesaurus of Psychological Index Terms

Contains over 4,000 terms representing an authoritative, unambiguous, and hierarchically structured vocabulary used by psychology and related disciplines. This vocabulary will facilitate precise content representation of concepts described in psychological literature and allow for more flexibility in automated search and retrieval operations. These two factors promote more effective communication of psychological concepts in original works and research, ultimately aiding the content of publications to be focused in accordance with the interest patterns of psychologists.

This **Thesaurus,** developed by the Psychological Abstracts Information Services, is divided into three sections:

The **relationship section** shows the relationships among concepts that are associated with each other. Term relationships noted are: synonymous, broader, narrower, or related.

The **alphabetical section** presents each preferred term in alphabetical order, enabling rapid identification of terms used for indexing.

The **hierarchical section** contains 17 broad content areas in which terms are presented in descending order according to breadth of concept represented. The 17 categories correspond to major classifications used in *Psychological Abstracts*.

The relationship, alphabetical, and hierarchical sections will assist the user in selecting appropriate terms for different uses.

Administrators
- determine areas of research for an organization
- assure representative expertise in personnel selection

Researchers and Authors
- find synonyms for a term representing a concept
- select terms that represent a concept
- define subject matter topics for a paper or book

Educators
- structure a curriculum
- use as a tutorial aid

Consultants
- search the literature
- aid in retrieval of information

Librarians
- provide access to psychological literature
- aid in cataloging psychology material
- aid in ordering new psychological literature

Students
- locate concepts relevant to psychology
- focus areas of study

The **Thesaurus** contains 362 pages with a hardcover and costs $12. Please note that all orders amounting to $15 or less must be accompanied by full payment except for those orders submitted on institutional purchase order forms.

American Psychological Association
Publication Sales Department
1200 17th Street, N.W.
Washington, D.C. 20036

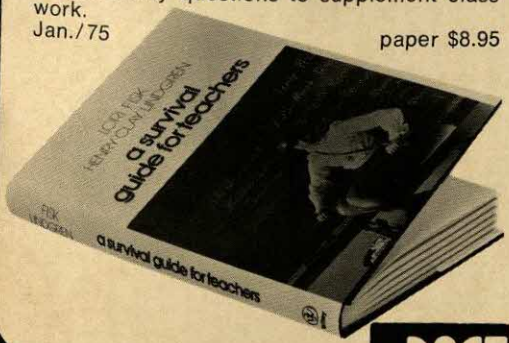
Tactical books that make tomorrow's decisions easier today

CONTEMPORARY EDUCATIONAL PSYCHOLOGY: Concepts, Issues, Applications
By Robert C. Craig, William A. Mehrens, and Harvey F. Clarizio, all of Michigan State University

This book provides future teachers with all the practical skills, current theory, and research they need to make a wide range of decisions about educational and social issues. Both readable and scholarly, it considers diverse psychological viewpoints and methods. It is divided into 5 sections, representing major areas in which teachers must make decisions: 1) overview and goals; 2) readiness in developmental aspects; 3) learning and instruction; 4) classroom dynamics; and 5) statistics, measurement, and evaluation. Each area is examined in terms of its application to the classroom and society. It also delves into the controversial issues confronting the profession, including: accountability, behavioral objectives, alternative schools, compensatory programs, and the school's role in reducing prejudice. Strategies and techniques essential to teaching success are featured, with emphasis on how to:

- organize learning material that provides reinforcement and stimulates feedback
- compare and combine divergent methods in problem solving
- select and use standardized tests

An Instructor's Manual keys essay and objective tests to chapter objectives and contains related study questions to supplement class work.
Jan./75 paper $8.95



A SURVIVAL GUIDE FOR TEACHERS
By Lori Fisk, Ralph A. Gates School, El Toro, California and Henry Clay Lindgren, California State University

This book tells you practical things about children in the classroom that most standard elementary texts let you discover on your own. It examines common experiences that prompt that "What do I do now?" feeling. Written by an experienced teacher/psychologist team, it is designed especially for the new teacher. Instead of chapters on curriculum change and public accountability, this book focuses on games children play, the first day of school and how to handle it, routine misbehavior, and how to get over wanting them to love you. It presents specific situations drawn from the authors' own classroom experiences with specific answers that show how to deal effectively with the day-today problems of classroom control, management and learning. Entertaining and instructive, this book will help support you in your first few teaching years "on the other side of the desk." 1973
176 pages cloth $9.75 paper $4.95

TEACHING OUR CHILDREN
By Joel E. Macht, University of Denver
This book is a must for teachers, students, parents, and future parents. It offers you ideas about how to teach children many kinds of desired behavior and helps children bring about some behavior they desire from you.

It focuses on everyday interactions between parents and children and the types of effects parents and teachers have on children's actions. You'll learn how to avoid teaching undesired effects and how to help children learn things which are most beneficial to them.
Jan./75

For your complimentary examination copies, contact your Wiley representative or write Richard Baker, Dept. 260, N.Y. office. Please include course title, enrollment, and present text.

JOHN WILEY AND SONS, INC.
605 Third Avenue, New York, N.Y. 10016
In Canada: 22 Worcester Road, Rexdale, Ontario
Price subject to change without notice.

Wiley

CHARLES C THOMAS • PUBLISHER

PSYCHO-EDUCATIONAL DIAGNOSIS OF EXCEPTIONAL CHILDREN *by* **Milton V. Wisland,** *Bellingham Public Schools, Bellingham, Washington. (9 Contributors)* Of interest to students of psychology, sociology and education, this text will also serve as an introduction to professional training programs dealing with exceptional children as well as a source of specific information for professionals. Students majoring in special education or clinical and school psychology will find this text a beneficial reference in their programs also. Three introductory chapters provide background for evaluating psycho-educational instruments, and the remaining seven chapters discuss specific assessment methods used with various exceptional children such as the deaf or gifted. Professional perspective can be broadened since the exceptional child is seen as the cooperative responsibility of many disciplines. *'74, 375 pp., 26 il., 27 tables, $13.95*

EFFECTIVE TEACHING METHODS FOR AUTISTIC CHILDREN *by* **Rosalind Oppenheim,** *The Rimland School for Autistic Children, Evanston, Illinois.* This book describes the teaching techniques she developed during her more than thirteen years in the field. However, this book is more than a mere description of techniques. It also outlines classroom approaches and child management methods that are designed to overcome the children's resistance and behavioral difficulties, and thus facilitate the learning process. Mrs. Oppenheim is a strong advocate of the idea that an effort should be made to teach academic readiness and the reading, writing and arithmetic to all autistic and autistic-type children, regardless of the abnormalities and developmental immaturities of their functioning. This book will prove valuable for both parents and professionals who are involved in the teaching of autistic children. Perhaps its most significant contribution is that it demolishes unequivocally the myth that autistic youngsters are unreachable and unteachable. *'74, 136 pp., $8.50*

FREUD AND EDUCATION *by* **Seymour Fox,** *the Hebrew Univ. of Jerusalem, Israel.* This book has been written for educators, psychologists and others interested in gaining an understanding of the structure of Freud's personality theory and its uses for education. A broader purpose is to offer a critical tool with which to examine the contributions different personality theories can make to educational theory and practice. Most of this book is devoted to disclosing the basic principles, terms and concepts of Freud's theory; and relating these to the issues that every personality theory must treat. In addition, textual analyses of four major theoretical works are appended in the form of protocols. Here analysis is joined with the author's comments, questions and disclosure of choice points in the development of the theory. The evidence for this interpretation is drawn from all of Freud's theoretical works. *'74, about 262 pp.*

PRESCRIPTIONS FOR CHILDREN WITH LEARNING AND ADJUSTMENT PROBLEMS (2nd Ptg.) *by* **Ralph F. Blanco,** *Temple Univ., Philadelphia.* This book gives specific help to practitioners in formulating psychological and educational recommendations after diagnostic assessment. Emphasis is given to learning and behavior problems in school and at home. Concepts are derived from psychodynamic, reinforcement and need theory orientations. Techniques, curricular suggestions and remedial ideas are given on withdrawn and aggressive behaviors; stealing, negativism, suicidal and homicidal behaviors; the blind, deaf and brain injured; retardation, the slow learner and learning disabilities; passive-aggressive behavior, underachievement, school phobia and many others. Over 3,700 recommendations from 146 experienced psychologists were edited after survey research to develop this comprehensive manual. Guidelines are offered to bridge the gap between diagnosis and treatment through an organization of possible solutions. *'74, 320 pp., $9.25*

Orders with remittance sent, on approval, postpaid

301-327 East Lawrence Avenue • Springfield • Illinois • 62717

Journal of Educational Psychology

VOLUME 66 NUMBER 6

DECEMBER 1974

AMERICAN PSYCHOLOGICAL ASSOCIATION

EDITOR

JOANNA WILLIAMS, *Teachers College, Columbia University and University of Pennsylvania*

ASSOCIATE EDITOR

ADVISORY EDITORS

PAUL R. AMMON, *University of California, Berkeley*
JERE BROPHY, *University of Texas at Austin*
C. VICTOR BUNDERSON, *Brigham Young University*
LEONARD S. CAHEN, *Educational Testing Service and Far West Laboratory for Educational Research and Development*
LOIS-ELLIN DATTA, *National Institute of Education*
FREDERICK B. DAVIS, *University of Pennsylvania*
FRANCIS J. DIVESTA, *Pennsylvania State University*
ROBERT EBEL, *Michigan State University*
DAVID ELKIND, *University of Rochester*
LAWRENCE T. FRASE, *Bell Laboratories*
N. L. GAGE, *Stanford University*
ROBERT GAGNÉ, *Florida State University*
ROBERT GLASER, *University of Pittsburgh*
MIRIAM L. GOLDBERG, *Teachers College, Columbia University*
EDMUND W. GORDON, *Teachers College, Columbia University*
VERNON C. HALL, *Syracuse University*
ARTHUR R. JENSEN, *University of California, Berkeley*
PHYLLIS A. KATZ, *City University of New York, Graduate School and University Center*
HERBERT J. KLAUSMEIER, *University of Wisconsin—Madison*
ROBERT E. LANA, *Temple University*
GERALD S. LESSER, *Harvard University*
JOEL R. LEVIN, *University of Wisconsin—Madison*
FREDERICK J. MCDONALD, *Educational Testing Service*
WILBERT J. MCKEACHIE, *University of Michigan*
FRANK B. MURRAY, *University of Delaware*
LAUREN B. RESNICK, *University of Pittsburgh*
WILLIAM D. ROHWER, JR., *University of California, Berkeley*
ROBERT ROSENTHAL, *Harvard University*
ERNST Z. ROTHKOPF, *Bell Laboratories*
S. JAY SAMUELS, *University of Minnesota*
SEYMOUR B. SARASON, *Yale University*
MELVYN I. SEMMEL, *Indiana University*
IRVING SIGEL, *Educational Testing Service*
HARRY SILBERMAN, *University of California, Los Angeles*
RICHARD E. SNOW, *Stanford University*
GEORGE G. THOMPSON, *Ohio State University*

JOURNAL STAFF

HAROLD P. VAN COTT, *Managing Editor*
ANITA DEVIVO, *Executive Editor and General Manager*
LINDA B. BARODY, *Supervising Technical Editor*
SHERRI RICARDO, *Technical Editor*
LYNN D. SPIER, *Acting Manager, Publications Business Services*
ANNE REDMAN, *Subscription Manager*

Published bimonthly in one volume per year, six issues (February, April, June, August, October, and December), by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D. C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D. C., and additional mailing offices.

The **Journal of Educational Psycholog** lishes original investigations and theoretical dealing with learning and cognition, especially relate to problems of instruction, and with the logical development, relationships, and adju of the individual. Preference is given to studie more complex types of behavior, especially in lating to educational settings. Journal articles to all levels of education and to all age groups.

Manuscripts: Submit manuscripts in triplicate Editor, Joanna Williams, P.O. Box 51, Teache lege, Columbia University, New York, New 10027, according to instructions provided else in the Journal.

Copyright and Permissions: All rights rese Written permission must be obtained from the A ican Psychological Association for copying or r ing text of more than 500 words, tables, or Permissions are normally granted contingent like permission of the author, inclusion of the copyright notice on the first page of repro material, and payment of $10 per page of APA rial. Permission fees are waived for authors wh to reproduce their own articles, or when mate reproduced in limited quantities for instruct purposes. Address requests for permission t Permissions Office, American Psychological As tion, 1200 Seventeenth Street, N. W., Washir D.C. 20036.

Subscriptions: Subscriptions are available calendar year basis only (January through Dece Nonmember rates for 1974: $24 domestic, $2 eign, $5 single issue. APA member rate: $10. W Subscription Section, APA.

Reprints: Authors may order reprints of their a from the printer when they receive galley proofs.

Back Issues and Back Volumes: For inform regarding back issues or back volumes write t der Department, APA.

Microform Editions: For information regard croform editions write to any of the following: Associates, Inc., P. O. Box 1017, Greenwich, ticut 06830; University Microfilms, Ann Arbor gan 48106; or Princeton Microfilms, Princeto Jersey 08540.

Change of Address: Send change of address and a recent mailing label to the attention of the scription Section, APA, 30 days prior to the change of address. Undelivered copies resulting address changes will not be replaced; second- mail is not forwarded unless subscribers notify local post office in writing that they will guara second-class forwarding postage.

JOURNAL OF EDUCATIONAL PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

ne 66 December 1974 Number 6

Notice to APA Authors 801

Long-Term Predictive Validity of Divergent-Thinking Tests: Some Negative Evidence 802
NATHAN KOGAN AND ETHEL PANKOVE

Influence of Children's Sex Role Standards on Reading and Arithmetic Achievement 811
CAROL ANNE DWYER

Children's Solution Processes in Arithmetic Word Problems 817
DANIEL J. A. ROSENTHAL AND LAUREN B. RESNICK

Relationship Between Mental Abilities, Social Class, and Exposure to English in Chinese Fourth Graders 826
LELAND Y. YEE AND ROLFE LAFORGE

Effect of Hypothesis/Test Training on Reading Skill 835
S. JAY SAMUELS, PATRICIA DAHL, AND TEARA ARCHWAMETY

Self-Determination of Academic Standards by Children: Toward Freedom from External Control 845
JEFFREY J. FELIXBROD AND K. DANIEL O'LEARY

Effect of Prior Teaching Evaluations and Lecture Presentation on Ratings of Teaching Performance 851
RAYMOND P. PERRY, R. RONALD NIEMI, AND KEITH JONES

Cognitive Results Based on Different Ages of Entry to School: A Comparative Study 857
GILBERT R. AUSTIN AND T. NEVILLE POSTLETHWAITE

Lesson Kinetic Structure Analysis as Related to Pupil Awareness and Achievement 864
ROBERT J. BROWNE AND O. ROGER ANDERSON

Effect of Two Free-Time Reinforcement Procedures on Academic Performance in a Class of Behavior Problem Children 872
DAVID MARHOLIN II, ELIZABETH T. MCINNIS, AND TOM B. HEADS

Retention of Pictorial and Verbal Content of A Text with Illustrations 880
J. PEECK

Serial Memory Span Thresholds of Normal and Mentally Retarded Children 889
ALFRED A. BAUMEISTER

Reading Grade Level Adjustment and Auditory Supplementation as Techniques for Increasing Textual Comprehensibility 895
ARTHUR I. SIEGEL, MARTIN R. LAUTMAN, AND JAMES R. BURKETT

Information Seeking Following the Confirmation or Contradiction of Beliefs 903
CHARLES B. SCHULTZ

Influence of Cultural Background on Hierarchical Learning 911
RUSSELL D. LINKE

Outerdirectedness in Children of Three Ages as a Function of Experimentally Induced Success and Failure 919
DONALD L. MACMILLAN AND DEBORAH L. WRIGHT

Diary No. 878
Date 30 4 | 75

(*Continued on next page*)

(*Continued from previous page*)

Training Letter Discrimination by Presentation of High-Confusion Versus Low-Confusion Alternatives 926
ROSEMERY O. NELSON AND KENNETH S. WEIN

Multitrait-Multimethod Analysis of Conceptual Tempo 932
VERNON C. HALL AND WILLIAM J. C. RUSSELL

Effects of Vicarious Consequences and Model and Experimenter Sex on Imitative Behavior in First-Grade Children 940
CANDACE S. GARRETT AND DONALD J. CUNNINGHAM

Learning Ability and Intelligence as Cognitive Predictors of Achievement in First-Grade Black Children 948
TREVOR E. SEWELL AND ROGER A. SEVERSON

Locus of Control and Aptitude Test Scores as Predictors of Academic Achievement 956
WALTER R. NORD, FRANCIS CONNELLY, AND GEORGE DAIGNAULT

An Exploration of Two Correction Procedures Used in Mastery Learning Approaches to Instruction 962
JAMES H. BLOCK AND MICHAEL L. TIERNEY

Locus of Control, Test Anxiety, and Student Performance in a Personalized Instruction Course 968
GEORGE J. ALLEN, LAURA GIAT, AND ROGER J. CHERNEY

Effects of Mode of Modeling, Model Age, and Ethnicity on Rule-Governed Language Behaviors 974
JAMES I. GRIESHOP AND MARY B. HARRIS

List of Manuscripts Accepted 910

This is the last issue of Volume 66.
Title page and Volume Contents appear herein.

SUGGESTIONS TO CONTRIBUTORS

In preparing articles for publication in the Journal, authors are strongly advised to follow the general directions given in the *Publication Manual of the American Psychological Association* (2nd ed.). All articles must be preceded by an abstract of 100–120 words. Instructions for preparation of the abstract appear in the Manual on p. 15. Special attention should also be given to the preparation of references (pp. 58–67, 122–130). In reference lists, give journal titles in full; do not abbreviate. *All copy must be double-spaced, including the references.* All manuscripts should be submitted in *triplicate*, one of which should be an original typed copy. Since the reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, the date the manuscript is submitted, and footnotes identifying the author or his affiliation. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Every effort should be made by the author to see that the manuscript itself contains no clues as to his identity. Dittoed and mimeographed copies are not acceptable and will not be considered. Original figures should be high-contrast, glossy, black and white prints of professional quality. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail and to check carefully the typing of the final copy.

Notice to APA Authors

The American Psychological Association announces publication of the second edition of the *Publication Manual* in August 1974. This new edition, which supersedes the 1967 Manual, will be adopted by the 14 APA journals in 1975.

The new Manual is more comprehensive than the previous edition. It updates APA publication policies and procedures and incorporates changes in editorial practice since 1967. For instance, APA now sends many authors their copy-edited manuscripts for review before they are set into type, and some editors now use blind review procedures. The second edition also includes up-to-date statements of the coverage of each APA journal including the *Journal of Experimental Psychology* which will be published in four separate sections in 1975.

The new *Publication Manual* initiates several changes in APA style. These changes are announced in the August 1974 *American Psychologist* and will be introduced in the APA journals in January 1975. During the period of transition to the new style, authors should note that (*a*) all manuscripts published in 1974 will be copy-edited according to the 1967 Manual, (*b*) manuscripts accepted in 1974 and published in 1975 will be copy-edited to conform to the new Manual. Starting in 1975, accepted manuscripts that depart significantly from the Manual will be returned to authors for correction.

Authors will be encouraged by the changes in the second edition. The new APA style simplifies reference forms; eliminates unnecessary underlines, brackets, and other devices; supports appropriate use of "I" and "we"; and generally clarifies typing requirements. Material is arranged for maximum convenience to authors and typists, and all sections are cross-referenced and indexed.

The new *Publication Manual* is available after August 1 for $3. Send orders to APA Publication Sales, 1200 Seventeenth Street, N.W., Washington, D.C. 20036. Orders of $15 or less must include payment unless they are submitted on institutional purchase order forms.

Journal of Educational Psychology
1974, Vol. 66, No. 6, 802-810

LONG-TERM PREDICTIVE VALIDITY OF DIVERGENT-THINKING TESTS: SOME NEGATIVE EVIDENCE[1]

NATHAN KOGAN[2]
New School for Social Research
ETHEL PANKOVE
Public Schools, Montgomery Township, New Jersey

Studied the relationship of assessments of divergent-thinking performance and intellective aptitude at fifth and tenth grades to nonacademic attainments assessed with a biographical questionnaire in the tenth grade and upon graduation. Stepwise multiple regression analysis based on middle-class subjects drawn from two school systems (*ns* of 46 and 22 after attrition) indicates that fifth-grade ideational productivity had no predictive power in respect to nonacademic attainment at graduation, whereas tenth-grade ideational productivity made a marginally significant contribution to that criterion in one of the school systems. Contrary to expectations, fifth- and tenth-grade intellective-aptitude measures account for modest to substantial amounts of the variation in nonacademic attainments. Assessment of such attainments is found to be fairly stable from the tenth grade to graduation. Examination of separate fields of nonacademic attainment indicates diverse relationships with cognitive assessments. Implications for creativity research are considered.

Creativity research based on the use of divergent-thinking tests has followed two relatively independent paths. The first of these can be subsumed under the general heading of construct validation and concerns the development of a theory to account for variation in divergent-thinking performance under natural conditions and as a consequence of diverse experimental treatments. Investigators in this tradition have examined cognitive and personality correlates of divergent-thinking performance and have constructed a variety of experimental procedures intended to enhance the level of ideational output on divergent-thinking tasks. Wallach (1970) has reviewed the foregoing areas of research and concluded that the ideational productivity or fluency component of divergent thinking is essentially independent of intelligence and reflects a process best characterized as "extensiveness of attention deployment."

In a more recent publication, Wallach (1971) has attacked the construct-validation approach to divergent thinking on the grounds that the search for correlates of ideational fluency and for methods of enhancing it has proceeded on the assumption of a powerful link between that form of

[1] This study was supported by the National Institute of Child Health and Human Development under Research Grant 5 P01 HD01762 to Educational Testing Service, Princeton, New Jersey. A number of individuals have assisted and participated in the project since its inception. We are especially grateful to Saul Cooperman and Edward McKeon of the Montgomery Township, New Jersey School System, and to Gary J. Estadt and Celeste Rorro of the Lawrence Township, New Jersey School System, for facilitating the research. We are also indebted to Edward Nystrom and Patricia Warren Cox for serving as experimenters, to Augusta Gross for coding of data, and to Henrietta Gallagher for supervision of data analysis. Thanks are due Norman Frederiksen and William C. Ward for comments and criticisms of an earlier version of the manuscript.

[2] Requests for reprints should be sent to Nathan Kogan, Department of Psychology, New School for Social Research, 66 West 12th Street, New York, New York 10011.

divergent thinking and genuine "real-world" creativity. Since the latter is obviously more difficult to assess than is ideational fluency, investigators have chosen to concentrate on such test-based measures with the justification that they are experimental analogues of "real-life" creative process. Such a stance prejudges the issue, of course, and constitutes the core of Wallach's criticism.

It is possible to maintain, then, that research on divergent thinking from a construct-validation perspective can be faulted in the absence of evidence for predictive or concurrent validation of test performance against "real-world" criteria. Such validation efforts constitute the second path that research on divergent thinking has taken. Torrance (e.g., 1972) reports a variety of validational evidence for his tests of creativity, but other investigators have seriously questioned the positive character of that evidence (e.g., Crockenberg, 1972; Harvey, Hoffmeister, Coates, & White, 1970). Further, Wallach (1970) has documented the lack of independence between the Torrance Tests and IQ indices. Cropley's (1972) five-year longitudinal study of the predictive validity of divergent-thinking tests is based in part on the Torrance procedures, hence, further complicating interpretation of the ambiguous outcomes of that study.

The Wallach and Kogan (1965) divergent-thinking tasks have been shown to manifest independence of IQ in a wide range of studies, but the examination of the concurrent and predictive validity of these tasks has been quite limited. Wallach and Wing (1969) obtained concurrent validity for the Wallach-Kogan procedures in a sample of recently graduated high school students. Extracurricular activities and accomplishments comprised the criteria, and these were *not* predicted by the verbal and mathematical scores of the College Board's Scholastic Aptitude Test. The overall effects were not especially strong, however, and the mode of analysis employed made it impossible to assess the portions of the variance in the criteria attributable to the divergent-thinking and intellective indices.

An initial attempt to explore the predictive validity of the Wallach-Kogan tasks is reported by Kogan and Pankove (1972). Children whose divergent-thinking performance and IQ level had been measured in the fifth grade were reexamined on analogous measures five years later in the tenth grade. At the latter time, an assessment of extracurricular participation was also obtained. The findings were mixed, the divergent-thinking tasks exhibiting significant concurrent and predictive validity in the smaller, but not the larger, school system. Also of interest was the significant independent contribution of IQ to the variation in the criteria in the smaller school system, suggesting that extracurricular accomplishments may be a joint function of divergent and convergent modes of thinking. The negative results in the larger school system, however, raise doubts about the potential replicability of the findings in the smaller school system and prompt consideration of the possible unreliability of the criterion measurement. The tenth grade is rather early in the student's high school career, and it is entirely feasible that the student's record of extracurricular activity and accomplishment upon graduation from high school would represent a more solid index of an "intermediate" creativity criterion.

In brief, the present paper offers an extension of the Kogan and Pankove (1972) study. Students whose divergent-thinking and intellective-aptitude levels had been assessed at fifth and tenth grades were surveyed at graduation in respect to their extracurricular involvements during their entire high school career. A rank in class index was also available upon graduation. The predictive validity (over approximately a seven-year and a two-year period) of divergent- and convergent-thinking tests in respect to "creative" and academic achievement is reported in the body of the paper.

Method

The fifth- and tenth-grade administrations of the Wallach-Kogan divergent-thinking measures are described in full in Pankove and Kogan (1968) and Kogan and Pankove (1972), respectively. Information on intellective-aptitude levels was avail-

able in school records. The present research has employed the overall IQ derived from the California Test of Mental Maturity and the power index derived from the Differential Aptitude Tests at fifth and tenth grades, respectively.

The biographical inventory of extracurricular activities and accomplishments devised by Wallach and Wing (1969), which had been administered in the tenth grade, was readministered in the twelfth grade just prior to graduation in the smaller school system and just after graduation in the larger school system. The graduating classes consisted of 102 and 258 students in the smaller and larger school systems, respectively. Both school systems can be described as predominantly middle class.

In the smaller school system, the Wallach-Wing questionnaire was personally given to all of the graduating seniors in the longitudinal sample by the school psychologist, and all of the questionnaires were returned. Nevertheless, considerable attrition occurred between fifth and twelfth grades. Out of a total of 46 children (25 males and 21 females) assessed in the fifth grade, a grand total of 22 (12 males and 10 females) graduated from the same school system seven years later.

In the larger school system, it was not possible to obtain the cooperation of all of the graduating seniors in the longitudinal sample. The Wallach-Wing questionnaire was mailed to all of the foregoing students with a covering note explaining the questionnaire's purpose and offering compensation of $2 for the return of the questionnaire in an enclosed self-addressed stamped envelope. A follow-up letter was sent to those students who had not returned the questionnaire within three weeks of the original mailing date. Of the 69 students who were mailed questionnaires, 46 (67%) returned them. Hence, of a total of 116 children (59 males and 57 females) examined in the fifth grade, only 46 (24 males and 22 females) were accessible for study seven years later.

In the Kogan and Pankove (1972) five-year follow-up, considerable attrition had already taken place—a 37.5% decline from the original sample size. Virtually all of this attrition could be attributed to family mobility. Those subjects who dropped out of the study between the fifth and tenth grades did not differ significantly in mean divergent-thinking performance or intellective aptitude from those subjects who were in the same school systems five years later.

In the case of the present seven-year follow-up, attrition is multiply determined—family mobility, dropping out of school prior to graduation, and failing to return the questionnaire. While there were no significant mean differences in fifth- and tenth-grade scores between those individuals who left the smaller school system between tenth and twelfth grades and those who remained, the same cannot be said for the larger school system. In the latter, those individuals who dropped out of the study between tenth grade and graduation (the large majority for failing to return the questionnaire) had significantly lower mean fifth- and tenth-grade intellective-aptitude scores than those who remained in the study (ts of 3.27 and 3.40, $p < .01$, respectively). The dropped subjects relative to those remaining in the study also scored lower on fifth-grade ideational productivity ($t = 2.67$, $p < .01$), but no difference was observed in the case of tenth-grade ideational productivity ($t < 1.0$).

In the earlier published reports (Kogan & Pankove, 1972; Pankove & Kogan, 1968), the Wallach and Kogan divergent-thinking tasks were scored for both productivity (number of responses) and uniqueness (number of unique responses). These two variables are so highly correlated, however, that only the ideational productivity data will be treated in the present report. In the earlier studies cited, within-task correlations between productivity and uniqueness ranged from .46 to .97 with a median correlation of .72.

On the Wallach-Wing biographical inventory, subjects' activities and accomplishments were scored in the areas of leadership, art, social service, literature, dramatic arts, music, and science. Each area was scored separately, and, in addition, a total score across all of the areas was obtained. Since the Wallach-Wing items within each area are largely cumulative (i.e., arranged from lowest to highest levels of accomplishment), weighted scores were used. For example, a subject who checked all of the items under one of the extracurricular areas would receive the same score as the subject who checked only that item representing the highest level of accomplishment.[3]

Results

Overall Extracurricular Activities and Accomplishments

Examination of the predictive validity of fifth- and tenth-grade productivity and intellective-aptitude measures in respect to total extracurricular activities and accomplishments over the high school years was carried out by means of stepwise multiple

[3] The quantification described does not, of course, provide a clear-cut discrimination between extracurricular activity, on the one hand, and genuine accomplishment, on the other. At best, we can merely state that those with high scores on the Wallach-Wing scales are higher on a combined activity-attainment index than those with low scores. Hence, when the terms "activity," "attainment," and "accomplishment" are employed in isolation in the text and tables, the combined index is always implied. The analytic separation of sheer extracurricular activity from genuine nonacademic accomplishment at the secondary school level constitutes an important problem in its own right but one that is beyond the scope of the present investigation.

TABLE 1

STEPWISE MULTIPLE REGRESSION ANALYSIS FOR PREDICTION OF OVERALL HIGH SCHOOL ACTIVITIES AND ACCOMPLISHMENTS

Predictor	Smaller school system				Larger school system			
	df	*R*	R^2	*F*	*df*	*R*	R^2	*F*
5th grade								
Productivity	20	−.038	.001	< 1.00	44	.012	.000	< 1.00
Intellective-aptitude	19	.673	.453	15.67***	43	.165	.027	< 1.00
10th grade								
Productivity	18	.741	.550	3.88*	42	.171	.029	< 1.00
Intellective-aptitude	17	.743	.551	< 1.00	41	.422	.178	7.42***
Activities	16	.786	.618	2.79	40	.699	.489	24.99***
Sex	15	.786	.618	< 1.00	39	.739	.545	4.82**

* $p < .10$.
** $p < .05$.
*** $p < .01$.

regression analysis. Tenth-grade activities and accomplishments and subject sex were also included as predictors in the analysis. Since the major aim of the research concerned long-term prediction as well as the specific efficacy of divergent-thinking tests, fifth-grade independent variables were entered in the analysis prior to the corresponding tenth-grade measures, and the divergent-thinking scores were considered prior to the intellective-aptitude measures. The outcomes for the smaller and the larger school systems are shown in Table 1.

It can be readily seen that ideational productivity assessed in the fifth grade completely fails to predict overall extracurricular activities and accomplishments assessed seven years later at graduation. This negative outcome holds for both school systems. In all other respects, the pattern of findings differs across the smaller and larger systems. In the former, intellective-aptitude level assessed at fifth grade is a powerful predictor of involvement and accomplishment in extracurricular activities several years later. Approximately 45% of the variance in such activities can be accounted for on the basis of subjects' intellective-aptitude level assessed in the later elementary school years. Only ideational productivity measured in the tenth grade contributes a marginally significant increment of predictive power, accounting for an additional 10% of the variance in the criterion. The negligible contribution of the tenth-grade activities index simply reflects the fact that fifth-grade intellective-aptitude level is almost as strongly correlated with it ($r = .52$) as with the criterion index ($r = .66$). In the Kogan and Pankove (1972) report, fifth-grade intellective-aptitude was a strong predictor of tenth-grade activities in the smaller school system, but both fifth- and tenth-grade productivity also accounted for significant portions of the criterion variance. The passage of time, then, appears to have had the effect of reducing the importance of the ideational productivity dimension and correspondingly enhancing the importance of intellective aptitude assessed as early as elementary school.

In the case of the larger school system, Kogan and Pankove (1972) observed that the combined fifth- and tenth-grade divergent-thinking and intellective-aptitude indices accounted for less than 5% of the variance in tenth-grade activities and accomplishments. Approximately two and a half years later, the picture has changed to some degree. Though fifth-grade measures continue to have virtually no predictive power, Table 1 indicates that tenth-grade intellective-aptitude level now makes a significant contribution, with total accountable variance in the criterion increased to approximately 18%. At the same time, tenth-grade activities clearly represent the strongest predictor, accounting for an additional 30% of criterion variance. In short, prior

cognitive assessments are less predictive of subsequent talented behavior in the larger than in the smaller school system. Earlier behavior of a similar nature is evidently the best predictor in the larger school system of the overall level of extracurricular involvement and accomplishment upon graduation. It should be noted, however, that the predictive power of the tenth-grade activities index is somewhat inflated by the part–whole relation to twelfth-grade activities. The latter index reflects total extracurricular performance and, hence, necessarily incorporates activities assessed two and a half years earlier. The significant F value for sex in the larger school system indicates a somewhat higher level of extracurricular activity and accomplishment for females.

Academic Achievement

The prediction of rank in class shows considerable correspondence between the two school systems (Table 2). In both, fifth-grade intellective aptitude makes a modest contribution to rank in class at graduation. Similarly, tenth-grade intellective aptitude in both school systems has considerable predictive power in respect to academic achievement, as indexed by rank in class. It should also be noted that apart from the sex variable in the larger school system (females manifest higher academic achievement), the accountable portion of the variance in the criterion is largely attributable to the cognitive assessments (primarily intellective aptitude). Tenth-grade activities and accomplishments do not offer an additional increment to the criterion variance in either school system. This latter finding supports the observations of Holland and Richards (1965) regarding the independence of academic and nonacademic accomplishment. Those authors also report that intellective-aptitude measures and nonacademic attainments at the high school level are independent in concurrent assessments. This kind of outcome was not obtained in the longitudinal assessments of the present study.

Activities and Accomplishments within Specific Fields

The relatively poor predictive power of the divergent-thinking measures in respect to overall nonacademic attainments may possibly be attributable to the unwarranted combination of such attainments across fields to yield a single total score. Though Wallach and Wing (1969) obtained evidence for concurrent validation using such a score, further analysis of their data indicated that the ideational productivity measures discriminated only for the areas of art, writing, science, and leadership. No such discrimination was obtained in the areas of dramatics, music, and social service. It is clearly important, then, to determine whether consideration of the particular areas of nonacademic attainment will fundamentally alter the form of the outcomes reported in Table 1.

Tables 3 and 4 contain the correlations

TABLE 2

STEPWISE MULTIPLE REGRESSION ANALYSIS FOR PREDICTION OF HIGH SCHOOL RANK IN CLASS

Predictor	Smaller school system				Larger school system			
	df	*R*	R^2	*F*	*df*	*R*	R^2	*F*
5th grade								
Productivity	20	.083	.007	< 1.00	44	−.070	.005	< 1.00
Intellective-aptitude	19	.397	.157	3.39*	43	.414	.172	8.26**
10th grade								
Productivity	18	.553	.306	3.86*	42	.471	.222	2.56
Intellective-aptitude	17	.825	.681	19.97**	41	.616	.380	9.93**
Activities	16	.827	.684	< 1.00	40	.625	.390	< 1.00
Sex	15	.838	.702	< 1.00	39	.707	.500	8.16**

* $p < .10$.
** $p < .01$.

TABLE 3

CORRELATIONS FOR THE SMALLER SCHOOL SYSTEM BETWEEN COGNITIVE ASSESSMENTS AND NONACADEMIC ATTAINMENTS IN DIFFERENT FIELDS

Predictor	Fields of nonacademic attainment							
	Leadership	Art	Social service	Writing	Dramatic arts	Music	Science	Total
5th grade								
Productivity	−.03	−.02	.05	−.11	.15	.03	−.18	−.04
Intellective-aptitude	.48*	.46*	.35	.47*	.28	.38	.44*	.66**
10th grade								
Productivity	.28	.27	.40	.27	.39	.48*	.01	.49*
Intellective-aptitude	.13	.20	.13	.21	.17	.38	.26	.36
Activity[a]	.60**	.49*	.58**	.48*	.56**	.62**	.05	.64**
Sex[b]	−.18	−.16	.08	.06	−.30	−.54*	.40	−.14
R	.74*	.63	.70	.63	.71	.83**	.65	.79**

Note. For the smaller school system, $n = 22$.

[a] This row presents correlations between the identical activity in 10th and 12th grades.

[b] Positive correlations reflect greater attainment by males, negative correlations by females.

* $p < .05$.

** $p < .01$.

between the cognitive assessments and nonacademic attainments in the separate fields for the smaller and larger school systems, respectively. Those tables also report the tenth- versus the twelfth-grade correlations for each field of nonacademic attainment, the correlation between sex and fields of nonacademic attainment, the multiple correlation for each field, and the correlations between each predictor and total nonacademic attainment.

A scan of the first row of Tables 3 and 4 quickly reveals the lack of any relationship of fifth-grade ideational productivity to nonacademic attainment in all of the fields surveyed. Hence, the absence of any predictive power for fifth-grade productivity observed in Table 1 is evidently not concealing any particular within-field effects. In contrast, fifth-grade intellective-aptitude level is significantly related to nonacademic attainment in several fields—leadership and

TABLE 4

CORRELATIONS FOR THE LARGER SCHOOL SYSTEM BETWEEN COGNITIVE ASSESSMENTS AND NONACADEMIC ATTAINMENTS IN DIFFERENT FIELDS

Predictor	Fields of nonacademic attainment							
	Leadership	Art	Social service	Writing	Dramatic arts	Music	Science	Total
5th grade								
Productivity	.07	−.14	.01	.08	.09	−.03	.00	.01
Intellective-aptitude	.35*	−.24	−.09	.29*	−.03	.16	.12	.16
10th grade								
Productivity	.17	.20	−.17	.22	.05	−.12	−.23	.01
Intellective-aptitude	.45**	.10	−.03	.35*	.09	.33*	.08	.39**
Activity[a]	.48**	.48**	.30*	.34*	.36*	.69**	−.12	.63**
Sex[b]	−.30*	−.18	.07	−.10	−.25	−.23	−.01	−.29*
R	.67**	.61**	.39	.46	.47	.75**	.35	.74**

Note. For the larger school system, $n = 46$.

[a] This row presents correlations between the identical activity in 10th and 12th grades.

[b] Positive correlations reflect greater attainment by males, negative correlations by females.

* $p < .05$.

** $p < .01$.

writing in both school systems as well as art and science in the smaller school system.

At tenth-grade level, there is a trend toward significant relations between ideational productivity and nonacademic attainment in the smaller school system, but only the correlation for music attains statistical significance. In the larger school system, ideational productivity at tenth grade is no more predictive of nonacademic attainment than was ideational productivity at fifth grade. In regard to tenth-grade intellective-aptitude level, the resemblance to the corresponding fifth-grade pattern is quite strong in the larger school system. For the smaller school system, on the other hand, tenth-grade intellective aptitude is a considerably poorer predictor of nonacademic attainments than is the corresponding fifth-grade index. Indeed, with the one exception of science, the ideational productivity correlations uniformly exceed the intellective-aptitude correlations in the tenth-grade assessment. This constitutes the sole evidence in the present study for the possible efficacy of ideational productivity for nonacademic attainments.

Discussion and Conclusions

On the whole, the results of the present longitudinal investigation support those authors who have questioned the practice of applying the "creativity" label to divergent-thinking tasks (e.g., Hudson, 1966; Nicholls, 1972). There is absolutely no indication in the data of the present study that divergent thinking assessed in the later years of elementary school is prognostic of nonacademic attainment during the high school years. It is possible to maintain, of course, that nonacademic attainment in secondary school represents a poor criterion of adolescent creativity. That may well be so, but it is difficult to envision a more desirable or appropriate one. It can also be argued that divergent-thinking tests might operate in terms of a "sleeper-effect" principle, lacking predictive validity in the high school adolescent years for "intermediate" criteria but gaining predictive power in respect to the "ultimate" criteria of mature adulthood. Again, such an effect may eventually be uncovered, but, until such time, the burden of proof clearly falls upon those who would equate performance on divergent-thinking tests with genuine creative achievement.

A lack of predictive validity does not necessarily imply the absence of concurrent or construct validity. The present investigation has merely shown that fifth-grade divergent-thinking performance is of little prognostic value in regard to talented nonacademic attainment in high school. There is good reason to believe that fifth-grade children with high scores on ideational productivity differ from those with low scores on a variety of cognitive and personality measures (Wallach & Kogan, 1965). The domain of childhood divergent-thinking assessment may be somewhat analogous to infant intelligence measurement, where it has also been demonstrated that infant test scores are highly informative of the child's cognitive status at the time but have little predictive validity for IQ at school age or later (e.g., Bayley, 1970).

Divergent-thinking measures obtained in tenth grade offered only a slight improvement over the comparable fifth-grade measures in respect to predictive validity. In the smaller school system, the former accounted for a marginally significant portion of the variance in nonacademic attainment at graduation. The previously reported (Kogan & Pankove, 1972) concurrent validity for tenth-grade ideational productivity was considerably more substantial in that school system, pointing to the weak "staying" power of such assessments.

The basis for the school system differences constitutes an intriguing unknown in the present research. In the absence of a thorough ecological analysis of the two school systems under study, we are obviously better equipped to raise questions than to provide answers. Given the evidence that 45% of the variation in nonacademic attainment at graduation in the smaller school system can be accounted for by an intelligence assessment made seven years earlier, in contrast to a mere 2.7% of

the variation in the larger system, one can conjecture as to whether the smaller school system subtly encourages its brighter youngsters to become involved in extracurricular activities, whereas such involvement develops more at the student's own initiative in the larger school system. Even at the level of tenth grade in the larger school system, intellective aptitude is accounting for less than 20% of the variance in nonacademic attainment. This leaves a great deal of room for the operation of noncognitive motivational and personality determinants. If a smaller school system encourages earlier and more extensive involvement in extracurricular pursuits, as proposed by Barker and Gump (1964) and empirically confirmed in the present samples (Kogan & Pankove, 1972), one might anticipate that the more popular, energetic, and uninhibited students would play a dominant role in such nonacademic activities. The modest predictive utility of tenth-grade ideational productivity hence suggests that the foregoing measure may be tapping dimensions of energy and responsiveness more than a cognitive capacity as such. Cronbach (1968) has offered a similar interpretation of ideational fluency.

The most consistent data in the present longitudinal study (apart from the negligible predictive validity of fifth-grade productivity measures) are the substantial associations between tenth- and twelfth-grade nonacademic attainments. Very likely, such associations will extend into the college years, as Richards, Holland, and Lutz (1967) have demonstrated. If future longitudinal research should demonstrate that these "intermediate" criteria are predictive of "ultimate" criteria of creative achievement in adulthood, we shall be forced to question the utility of cognitive testing for "creativity" in the high school years or even earlier (if reliable indices of nonacademic attainment are available). Such consistency cannot be taken for granted, however. The base rates for genuine adult creativity are likely to be exceedingly low, assuming that a generally acceptable operational definition of it can be formulated (see Kogan, 1973). Until research has clarified the issue of the link between high school or college "creativity" and subsequent adult creativity, the concern expressed in certain quarters (e.g., Wing & Wallach, 1971) about possible talent loss in the college admissions process may well be premature.

REFERENCES

Barker, R. G., & Gump, P. V. *Big school, small school.* Stanford, Calif.: Stanford University Press, 1964.

Bayley, N. Development of mental abilities. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology.* Vol. 1. New York: Wiley, 1970.

Crockenberg, S. B. Creativity tests: A boon or boondoggle for education. *Review of Educational Research,* 1972, **42,** 27–46.

Cronbach, L. J. Intelligence? Creativity? A parsimonious reinterpretation of the Wallach-Kogan data. *American Educational Research Journal,* 1968, **5,** 491–511.

Cropley, A. J. A five-year longitudinal study of the validity of creativity tests. *Developmental Psychology,* 1972, **6,** 119–124.

Harvey, O. J., Hoffmeister, J. K., Coates, C., & White, B. J. A partial evaluation of Torrance's Tests of Creativity. *American Educational Research Journal,* 1970, **7,** 359–372.

Holland, J. L., & Richards, J. M., Jr. Academic and nonacademic accomplishment: Correlated or uncorrelated? *Journal of Educational Psychology,* 1965, **56,** 165–174.

Hudson, L. *Contrary imaginations.* London: Methuen, 1966.

Kogan, N. Creativity and cognitive style: A life span perspective. In P. B. Baltes & K. W. Schaie (Eds.), *Life-span developmental psychology: Personality and socialization.* New York: Academic Press, 1973.

Kogan, N., & Pankove, E. Creative ability over a five-year span. *Child Development,* 1972, **43,** 427–442.

Nicholls, J. G. Creativity in the person who will never produce anything original and useful: The concept of creativity as a normally distributed trait. *American Psychologist,* 1972, **27,** 717–727.

Pankove, E., & Kogan, N. Creative ability and risk taking in elementary school children. *Journal of Personality,* 1968, **36,** 420–439.

Richards, J. M., Jr., Holland, J. L., & Lutz, S. W. Prediction of student accomplishment in college. *Journal of Educational Psychology,* 1967, **58,** 343–355.

Torrance, E. P. Predictive validity of the Torrance tests of Creative Thinking. *Journal of Creative Behavior,* 1972, **6,** 236–252.

Wallach, M. A. Creativity. In P. H. Mussen (Ed.),

Carmichael's manual of child psychology. Vol. 1. New York: Wiley, 1970.

Wallach, M. A. *The intelligence/creativity distinction*. New York: General Learning Press, 1971.

Wallach, M. A., & Kogan, N. *Modes of thinking in young children*. New York: Holt, Rinehart & Winston, 1965.

Wallach, M. A., & Wing, C. W., Jr. *The talented student: A validation of the creativity-intelligence distinction*. New York: Holt, Rinehart & Winston, 1969.

Wing, C. W., Jr., & Wallach, M. A. *College admissions and the psychology of talent*. New York: Holt, Rinehart & Winston, 1971.

(Received December 27, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 811-816

INFLUENCE OF CHILDREN'S SEX ROLE STANDARDS ON READING AND ARITHMETIC ACHIEVEMENT

CAROL ANNE DWYER[1]

Educational Testing Service, Princeton, New Jersey

The relationship between sex role standards and reading and arithmetic achievement was examined. It was hypothesized that children's sex role standards, assessed by checklist, would predict their achievement test scores. Subjects were 385 middle-class Caucasian children in Grades 2, 4, 6, 8, 10, and 12. Multiple regression analyses indicated that sex role standards contributed significant variance to reading and arithmetic achievement test scores. This effect was stronger for males than females. The results suggest that reading and arithmetic sex differences are more a function of the child's perception of these areas as sex-appropriate or sex-inappropriate than of the child's biological sex, individual preference for masculine or feminine sex role, or liking or disliking of reading or arithmetic.

This study examined the relationship between sex role standards (the extent to which the individual considers certain activities appropriate to males or to females) and achievement in the areas of reading and arithmetic. It was hypothesized that for both females and males, sex role standards would be a contributing factor to achievement and would predict achievement scores in reading and arithmetic less well than IQ but better than biological sex or liking or disliking the subject.

The concept of sex role standards was originally set forth by Kagan (1964), who found that second and third graders considered many school-related objects and activities feminine. This concept was substantially developed by Stein and Smithells (1968, 1969), Stein (1971), and Stein, Pohly, and Mueller (1971), who examined age and sex differences in children's sex role standards about achievement. Stein and Smithells (1969) found that both females and males considered reading activities feminine and arithmetic activities masculine.

Considerable research has been done on the subject of sex differences in reading and arithmetic achievement. In general, females have been found to excel in reading, and males in arithmetic, but the differences have varied with age, socioeconomic status, IQ, and the specific subskill being measured (Anastasi, 1958; Dwyer, 1973; Lewis, 1968; Maccoby, 1966). It was concluded for the purposes of this study that biological sex and liking or disliking a subject area were insufficient explanations for observed patterns of sex differences in achievement. Several studies have attempted to relate sex role standards to achievement behaviors. Carey (1958) showed that for male and female college students, sex differences in problem-solving proficiency were in part a reflection of attitudes toward the sex-appropriateness of problem solving and could be diminished through group discussions. Milton (1959) found that when the characteristics of problems were made less appropriate to the masculine sex role, the sex differences in problem-solving performance which had previously been observed were diminished.

Stein, Pohly, and Mueller (1971) found that two determinants of achievement motivation—attainment value and expectancy of success—were influenced by the individual's perception of the sex-role appropriate-

[1] Requests for reprints should be sent to Carol Anne Dwyer, Educational Testing Service, Princeton New Jersey 08540.

ness of the task. Subjects showed higher attainment value and expectancy of success on sex-appropriate tasks.

Only one other study (Mazurkiewicz, 1960) had dealt directly with the relationship between sex role standards and academic achievement. Using a sample of eleventh-grade males, Mazurkiewicz found that reading achievement test scores were higher for those boys who considered reading a masculine activity. This is the only study that has made an attempt to relate sex role standards to actual achievement test information, and even here, only males were used as subjects and only at one grade level.

Method

Subjects

The subjects were 385 Caucasian children attending public schools in a suburban, northern California community. All lived in a small geographical area consisting of one large, homogeneous housing development whose residents were all of similar socioeconomic status (predominantly high school graduates, employed in clerical capacities and as skilled or semiskilled workers). The subjects were approximately equally divided between males and females and approximately equally divided among Grade Levels 2, 4, 6, 8, 10, and 12.

The mean IQ for this sample was 105.8, based on group IQ tests routinely administered by the school district.

Materials

Information obtained for each subject included IQ scores; reading and arithmetic achievement test scores (the Stanford Achievement Test in Grades 2, 4, 6, and 8, and the Iowa Tests of Educational Development in Grades 10 and 12); a sex role standards checklist scored for both reading and arithmetic; and an individual sex role preference checklist scored for reading, arithmetic, and sex typing.

The sex role standards (SRS) checklist and the individual sex role preference (ISRP) checklist were modeled after checklists devised by Stein and Smithells (1969) to assess children's sex role standards about achievement and other areas. Additional items were written for the present study in the areas of reading and arithmetic, since Stein and Smithells' lists contain only six items in each of these categories. The SRS and ISRP checklists each consist of 46 items representing interests and activities in many areas, including 10 each in reading and arithmetic. The two checklists contained the same items but in different order and with different directions to the subject: The SRS checklist asked the subject which interests and activities he or she thought boys or girls preferred; the ISRP checklist asked the subject which interests and activities he/himself or she/herself preferred. On the SRS, the subject would respond by circling a "B" (boys prefer this activity), an "S" (both sexes like this activity equally), or a "G" (girls prefer this activity). On the ISRP checklist, the subject could respond by circling an "L" (I like this activity), a "?" (I cannot decide whether I like this activity), or a "D" (I dislike this activity).

Procedures

One female examiner administered all tests, with the exception of the tenth- and twelfth-grade achievement tests, which were administered by high school personnel of both sexes. Tests were administered in class-sized groups. In Grade 2, only the checklists were administered; this was done in small groups in which the examiner read each item aloud and the child marked his or her own response. The checklists were introduced informally as questionnaires designed to assess responses of students of different ages and sexes. The use of the "S" response in the SRS checklist and the "?" response in the ISRP checklist was mildly discouraged.

The SRS checklist was scored by assigning a score of 1 to the response "B," 2 to the response "S," and 3 to the response "G." The checklist originally contained 10 items representing reading and 10 representing arithmetic, but 2 of the 10 reading items and 3 of the 10 arithmetic items were eliminated because they failed to correlate highly enough with the total score for each category. The possible range of SRS scores, then, was from 8 (all 8 items judged masculine) to 24 (all 8 items judged feminine) for reading and from 7 (all 7 items judged masculine) to 21 (all 7 items judged feminine) for arithmetic.

Three scores were computed for the ISRP checklist: a reading preference score ($ISRP_r$), an arithmetic preference score ($ISRP_a$), and a sex role conformity (sex-typing) score ($ISRP_{st}$). The ISRP checklist was scored by assigning a score of 1 for each reading or arithmetic item disliked, 2 for each response of "?," and 3 for each item liked. The ISRP sex-typing score was based on the number of items the subject actually liked or disliked which had previously been chosen most often by the entire group as being appropriate for boys or for girls. That is, the total peer group responses on the SRS checklist determined which items (among all 46) were the most masculine and which were the most feminine, and the $ISRP_{st}$ score reflected the degree to which the individual subject conformed to or deviated from the items most strongly associated (by group consensus) with his or her sex. The total sex-typing score had a possible range of from 20 to 60 (in the former, the subject disliked all the

items chosen by his or her peers as being appropriate to his or her sex and liked all the items chosen by his or her peers as being inappropriate to his or her sex—the lowest sex-typed score—and in the latter, the subject liked all the items chosen by the peer group as being appropriate to his or her sex and disliked all the items chosen by the peer group as inappropriate to his or her sex—the most highly sex-typed score.)

Results

The central hypotheses of this study were tested by a series of multiple regression analyses, applications developed by Adler (1971) and Cohen (1968). Reading and arithmetic data were treated separately, first for the entire sample, then for each grade level separately, and then for males and females separately. In each of the multiple regression analyses, the relative importance of each variable in predicting achievement scores was considered in two ways. First was the unique variance, an estimate of the importance of a variable apart from the effects of all the other variables that were entered into the regression equation. This estimate is obtained by entering all the other variables except the specified variable into the regression equation and subtracting the resulting sum (the variance accounted for by all the other variables) from the estimate obtained by entering all the variables (including the specified variable) into the equation. The second estimate of the relative importance of a variable is the combination of a variable's unique variance it holds in common with the other variables.

Reading

Table 1 indicates that SRS contributed more unique plus common variance than sex or $ISRP_r$ and less unique plus common variance than IQ or grade level (Items A through E, Column 3). The SRS_r accounted for 4.5% of the variance in reading achievement scores when considered in combination with other variables. With the effects of the other variables partialed out, SRS_r accounted for 1.5% of the total variance. This figure is significant at the .001 level.

The sex-typing score derived from the ISRP checklist was found to be nearly unrelated to reading achievement scores. The unique plus common variance accounted for by $ISRP_{st}$ was .01%, and the unique variance accounted for was .0%.

The relationship of IQ, SRS_r, and sex to reading achievement scores was analyzed for the entire sample and for each grade level separately. In Grades 2 and 4, the unique variance accounted for was negligible (.6% and .01% respectively). The effects of SRS_r were strongest in Grades 8 and 10, accounting for 7.75% and 4.4%, respectively.

Data were also analyzed separately for females and for males. The effects of IQ on achievement were approximately equal for females and males. The effects of SRS_r were stronger for males than for females (for fe-

TABLE 1

Multiple Regression: IQ, SRS_r, $ISRP_r$, Sex, Grade Level, and Reading Achievement

Independent variable	R	R^2	Increment
IQ (A)	.4038	.1630	
SRS_r (B)	.2129	.0453	
$ISRP_r$ (C)	.0286	.0008	
Sex (D)	.0381	.0015	
Grade Level (E)	.3189	.1017	
A + B + C + D + E	.6104	.3726	
			A over (B + C + D + E) = .2171
			B over (A + C + D + E) = .0151*
			C over (A + B + D + E) = .0022
			D over (A + B + C + E) = .0034
			E over (A + B + C + D) = .1782

Note. Abbreviations: SRS_r = sex role standards checklist, reading; $ISRP_r$ = individual sex role preference checklist, reading.

* $p < .001$.

males .0198 unique variance contributed by SRS_r; for males, .0351 unique variance).

In sum, sex role standards about reading still accounted for between 1% and 8% of the total variance in reading achievement test scores even when the effects of sex, IQ, and liking or disliking reading were partialed out. The effect held true at all grade levels but was strongest at Grade Levels 6 and above.

Arithmetic

Table 2 describes the relationship of the variables IQ, SRS_a, $ISRP_a$, sex, and grade level to arithmetic achievement test scores for the entire sample. This table indicates that SRS_a contributed more unique plus common variance than $ISRP_a$ or sex and less unique plus common variance than grade level or IQ. The SRS_a accounted for 5.97% of the variance in arithmetic achievement scores when considered in combination with other variables. With the effects of the other variables partialed out, SRS_a contributed 2.6% unique variance. This figure is significant at the .01 level.

The sex-typing score derived from the ISRP checklist was found to be nearly unrelated to arithmetic achievement scores (as it was to reading achievement scores). The unique plus common variance accounted for by $ISRP_{st}$ was .006% and unique variance accounted for by $ISRP_{st}$ was .008%.

The relationship of IQ, SRS_a, and sex to arithmetic achievement scores was analyzed for the entire sample and for each grade level separately. The unique variance accounted for by SRS_a was lowest in Grade 12 (.3%) and highest in Grades 8 and 10 (5.67% and 8.7%, respectively). In this analysis by separate grade levels, it was only at Grades 8 and 10 that the SRS_a unique variance figure reached an acceptable level of significance.

Data were analyzed for the entire sample and for females and males separately. IQ contributed more unique variance to arithmetic achievement score for males than for females. There was no significant sex difference in the correlations of IQ with reading, but for females the correlation between IQ and reading was significantly higher than the correlation between IQ and arithmetic. As in reading, the effects of sex role standards in arithmetic were stronger for males than for females. For males the reverse was true. IQ was a better predictor of girls' reading achievement than their arithmetic achievement (and the reverse for boys), and it is assumed that social (sex role standards) factors accounted for part of this difference.

Sex role standards about arithmetic still accounted for between 1% and 9% of the variance in arithmetic achievement scores, even when the effects of IQ, sex, and liking or disliking arithmetic were controlled.

The relationship between sex role standards and arithmetic achievement held true

TABLE 2

MULTIPLE REGRESSION: IQ, SRS_a, $ISRP_a$, SEX, GRADE LEVEL, AND ARITHMETIC ACHIEVEMENT

Independent variance	R	R^2	Increment
IQ (A)	.3967	.1574	
SRS_a (B)	.2444	.0597	
$ISRP_a$ (C)	.0254	.0006	
Sex (D)	.0870	.0076	
Grade (E)	.3383	.1144	
A + B + C + D + E	.6026	.3631	
			A over (B + C + D + E) = .1902
			B over (A + C + D + E) = .0260*
			C over (A + B + D + E) = .0007
			D over (A + B + C + E) = .0068
			E over (A + B + C + D) = .1604

Note. Abbreviations: SRS_a = sex role standards checklist, arithmetic; $ISRP_a$ = individual sex role preference checklist, arithmetic.

* $p < .01$.

for both males and females but, as in reading, was stronger for males than for females (1% vs. 4%).

Contrary to expectation, individual sex role preference scores contributed very little variance to reading or arithmetic achievement. No explanations for this were apparent, but it was also noted that the expected sex differences in $ISRP_r$ and $ISRP_a$ did not emerge for this sample. $ISRP_r$ scores favored girls (girls stated that they liked reading more than boys so stated), but there were no significant sex differences in $ISRP_a$ scores.

Sex Role Standards

The SRS_r scores were significantly ($t = 12.08$, $df = 352$, $p < .01$) higher than 16, indicating that the subjects considered reading more feminine than masculine. Similar results were obtained for females and males at every grade level. The SRS_a scores were significantly lower than the expected mean value of 14 only for females in Grades 10 and 12 ($t = -2.16$, $df = 32$, $p < .05$; $t = -2.39$, $df = 27$, $p < .05$).

Other Sex Differences

To determine sex differences, t tests were carried out. There were no significant differences in ISRP sex-typing scores, in IQ, or achievement scores until Grade 12, where the difference favored males. Differences in arithmetic achievement scores favored boys at all grade levels but reached significance only in Grade 2 and for all grade levels combined.

All tests for form and order of administration effects were found to be nonsignificant.

Discussion

The stronger effect of sex role standards on males' than on females' achievement is most probably a reflection of the greater latitude our culture allows females in participating in the male role than males in participating in the female role.

The response patterns on the sex role standards checklist also support the observation that stronger social sanctions exist against males participating in the female sex role. The response patterns suggest that while males are subject to these greater sanctions, it is they themselves who define them. The response patterns on the sex role standards questionnaire showed that there was more agreement between the sexes as to what constituted girls' things than there was as to what constituted boys' things and that it was the boys who accounted for this disparity. Boys were more likely to label things as exclusively male than girls were. However, girls were *not* significantly more likely than boys to label things feminine (that is, boys *agree* with girls as to what constitutes feminine activities). Boys also gave fewer "S" responses than girls on the sex role standards checklist and more responses indicating that certain interests and activities are appropriate to their own sex.

Another interesting finding in the SRS checklist response patterns was a trend of the younger subjects in this sample to show more rigidity concerning sex roles than the older subjects. On the sex role standards checklist, older subjects made many more responses of "S" than did the younger subjects. This is in direct contradiction to the common research assumption that a high degree of sex role differentiation is a characteristic of increasing personal maturity. It is possible that this finding reflects a growing awareness of the personally limiting properties of strong sex role typing (as opposed to individual choice) and that a mature synthesis may give higher priority to personal abilities and interests than to maintenance of social sex role traditions.

References

Adler, P. T. Ethnic and socioeconomic status differences in human figure drawings. *Journal of Consulting and Clinical Psychology*, 1971, **36**, 344–354.

Anastasi, A. *Differential psychology*. (3rd ed.) New York: Macmillan, 1958.

Carey, G. Sex differences in problem solving as function of attitude differences. *Journal of Abnormal and Social Psychology*, 1958, **56**, 256–260.

Cohen, J. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 1968, **70**, 426–443.

Dwyer, C. A. Sex differences in reading: An evaluation and a critique of current theories. *Review of Educational Research,* 1973, **43,** 455–467.

Kagan, J. The child's sex role classification of school objects. *Child Development,* 1964, **35,** 1051–1056.

Lewis, E. C. *Developing woman's potential.* Ames: Iowa State University Press, 1968.

Maccoby, E. E. *The development of sex differences.* Stanford, Calif.: Stanford University Press, 1966.

Mazurkiewicz, A. J. Social-cultural influences and reading. *Journal of Developmental Reading,* 1960, **3,** 254–263.

Milton, G. A. Sex differences in problem solving as a function of role appropriateness of the problem content. *Psychological Reports,* 1959, **5,** 705–708.

Stein, A. H. The effects of sex-role standards for achievement and sex-role preference on three determinants of achievement motivation. *Developmental Psychology,* 1971, **4,** 219–231.

Stein, A. H., Pohly, S. R., & Mueller, E. The influence of masculine, feminine, and neutral tasks on children's achievement behavior, expectancies of success, and attainment values. *Child Development,* 1971, **42,** 195–207.

Stein, A. H., & Smithells, J. The sex-role standards about achievement held by Negro and White children from father-present and father-absent homes. Unpublished manuscript, Cornell University, 1968.

Stein, A. H., & Smithells, J. Age and sex differences in children's sex-role standards about achievement. *Developmental Psychology,* 1969, **1,** 252–259.

(Received December 26, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 817-825

CHILDREN'S SOLUTION PROCESSES IN ARITHMETIC WORD PROBLEMS[1]

DANIEL J. A. ROSENTHAL[2] AND LAUREN B. RESNICK

University of Pittsburgh

The influence of mathematical and linguistic factors in solving verbally presented arithmetic problems was investigated in two groups of third-grade children. Problems in which events were mentioned out of chronological order were more difficult to solve. Problems with the starting set unknown (verbal representations of equations such as $_ + c = e$) were more difficult and took longer than ending set unknown problems. An interaction indicated that problems with verbs denoting a gain (verbal representations of addition equations) were more difficult than loss verb problems when the starting set was unknown, and easier when the ending set was unknown. Two possible solution process models for unknown starting set problems are discussed on the basis of solution time data.

Word problems in arithmetic are tasks which require the integration of linguistic and arithmetic processing skills. In word problems, a situation is described in which there is some modification, exchange, or combination of quantities. On the basis of this verbal description, the individual solving the problem must construct a representation of the arithmetic operations called for; this representation mediates the arithmetic solution process.

There is a small body of research on arithmetic problems which attempts to uncover the processes involved in performing arithmetic operations such as addition and subtraction (cf. Groen & Parkman, 1973; Parkman & Groen, 1971; Restle, 1970; Suppes & Groen, 1967; Woods, 1972). In this research, problems are presented numerically in the form of simple algebraic equations. A few studies of this kind have related the position of the unknown set to the solution process employed (cf. Groen & Poll, 1973; Peterson & Aller, 1971; Suppes, Hyman, & Jerman, 1967), but in this work there has been no attention to the role of verbal or situational context on arithmetic performance.

Another small body of research on arithmetic word problems has attempted to relate language variables to the way in which verbal messages are encoded and stored. Steffe (1967) showed that children made fewer errors solving problems in which the names of the sets to be added were the same (e.g., cars and cars) than in problems where the set names were different (e.g., jacks and marbles). Loftus and Suppes (1972) found a significant relationship between the structural complexity of the language in a problem and the probability of making a correct response. Paige and Simon (1966) performed a detailed analysis of the way in which adults translate verbal information in algebra word problems into equations;

[1] The research reported herein was supported by funds from the Office of Naval Research, Contract N0014-67-A-0402-0006, and by the Learning Research and Development Center, University of Pittsburgh, supported in part as a research and development center by the United States Office of Education, Department of Health, Education and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of either organization, and no official endorsement should be inferred.

This research was conducted as a master's thesis by Daniel Rosenthal (1971) under the direction of Lauren Resnick.

[2] The authors wish to thank Guy Groen for his many helpful suggestions.

Requests for reprints should be sent to Daniel J. A. Rosenthal, who is now at the Department of Educational Psychology, College of Education, Temple University, Philadelphia, Pennsylvania 19122.

however, they did not attempt to study the effect of these representational processes on the actual solution of the equations.

Outside the arithmetic context, there exists a large and varied body of literature on verbal comprehension processes, that is, the ways in which verbal messages are represented and stored. In much of this research, the organization of recall is used as the index of representational strategy. There has been little exploration of the way in which different forms of verbal expression and representation might affect the use of information in subsequent problem-solving behavior.

The study of processes used in solving simple arithmetic word problems affords an opportunity to study the relationship between verbal processes and nonlinguistic task performance in a context that has a relatively high degree of "ecological validity." Most arithmetic processing tasks that are encountered outside the classroom arise from a situation in which equations or equivalent representations must be constructed as well as solved. Similarly, verbal information is typically recalled or reconstructed for use in some problem-solving or communication task, a condition difficult to simulate using typical laboratory tasks.

The present study focuses on a carefully defined set of word problems that permits examination of the effects of specific aspects of the verbal stimuli and the problem situation. All problems required only a single operation, either addition or subtraction. All quantities manipulated were greater than one and less than 10. All problems were presented in a verbal form that contained three separate clauses, one describing the "starting set," one describing the "change set," and one describing the "ending set." The problems differed in three dimensions: (*a*) the *order of mention* of chronological events, that is, the sets were mentioned in chronological order (starting, change, ending) or in reverse order (ending, change, starting); (*b*) the *identity of the unknown set,* that is, either the starting or the ending set was the unknown set in a problem (e.g., "5 + 2 = __" versus "__ + 2 = 5"); and (*c*) the *type of verb* associated with the change set; the verb expressed either a gain or a loss, suggesting an initial equation involving addition or subtraction, respectively.

Order of mention. Prior research has shown that when the order in which events are mentioned in a verbal expression does not match the chronological order of occurrence, recall is more difficult. Clark and Clark (1968) argued that events mentioned out of temporal order (as, for example, in the sentence, "He swiped the cabbages after he tooted the horn [p. 130]," are marked in memory so as to distinguish the order of occurrence from the order of mention. Bever (1970) suggested that underlying the Clarks' results was a basic habit of listening in which "in the comprehension of ordered events we organize relations by starting with the first event, organizing other events as subsidiary to the first [p. 286]." Huttenlocher and Weiner (1971) have shown that when ordering objects in response to a descriptive statement of their relationships, children have a tendency to move the first-mentioned object first. Thus, they begin by processing the first-mentioned item.

This research suggests that in a word problem in which order of mention is the reverse of chronological order, subjects' normal processing habits are disrupted. They may either treat the first-mentioned set erroneously as the starting set, that is, fail to mark the discrepancy of orders; or they may search for the starting set and then organize the remainder of the problem around it. Thus, the following two hypotheses can be made for the order of mention variable:

1. There are more errors in backward order of mention than in forward order of mention problems.
2. For correctly solved problems, backward order problems have longer solution times than forward order problems.

Identity of the unknown set. Work by Suppes et al. (1967) has established that for algebraically presented addition problems of the form s + c = e (where the symbols *s*, *c*, and *e*, refer to the starting set, change set, and ending set, respectively), problems where the starting set is unknown (i.e., __ + c = e) are more difficult and take longer to

solve than problems where the ending set is unknown (i.e., s + c = __). The increased difficulty of problems where the initial set is unknown presumably results from the different procedures required for the solution of problems in this form. Previous research with algebraically presented addition problems where the final set was the unknown, that is, s + c = __ (cf. Suppes & Groen, 1967; Groen & Parkman, 1972), has supported a counting model of problem solution in which children store the larger of the two given numbers in a mental counter, and then increment by the smaller number. Although no solution model for starting set unknown problems has been clearly supported (cf. Groen & Poll, 1973; Suppes et al., 1967), it is clear that the solution procedures for problems where the ending set is unknown cannot be directly applied to these problems because (*a*) the presence of a starting set is a requirement for this model to operate, and (*b*) in starting set unknown problems, the unknown set does not stand alone to one side of the equal sign, so the procedure of combining the two given numbers using the given operation does not give the correct answer. This means that some additional operations must be involved in solving unknown starting set problems. The following hypotheses can thus be stated for the identity of the unknown set variable:

3. Word problems where the unknown set is the starting set are more difficult to solve than problems where the unknown set is the ending set.

4. For correctly solved problems, unknown starting set problems have longer solution times than unknown ending set problems.

Type of verb. There is no basis in past research for making specific predictions concerning the effect of gain versus loss verbs. The relative difficulty of numerically presented simple addition and subtraction problems (i.e., s + c = __, and s − c = __) has never been directly compared, to our knowledge. Further, in our problems, gain verbs do not necessarily imply the use of addition operations nor do loss verbs necessarily imply the use of subtraction operations. Problems where the starting set in unknown, as in problems based on __ + c = e, might be solved by transformation into an equivalent canonical form, i.e., e − c = __ (cf. Suppes et al., 1967), thus, involving a subtraction operation in a gain verb problem. Alternatively, such problems might be solved by a trial and error procedure in which various values of s are tested directly in the initial __ + c = e equation, thus using an addition operation. The data from the present experiment should allow more specific hypotheses concerning solution of problems of these kinds to be formulated.

Method

Subjects

Two groups of third-grade children from two public schools served as subjects in two replications of the experiment. The subjects in both schools were from predominantly white working-class families. The 29 subjects in Experiment 1, 15 boys and 14 girls, were just beginning the third grade. The 34 subjects in Experiment 2, 15 boys and 19 girls, were just completing the third grade. Both subject groups represented a cross-section of the pupil population with regard to academic standing.

Stimulus Materials

Item forms (Hively, Patterson, & Page, 1968; Osburn, 1968) were used to generate a series of arithmetic word problems which systematically varied with respect to the three factors previously defined: order of mention, identity of the unknown set, and type of verb. Initially, we defined (*a*) a set of generalized frames or item forms, containing both fixed and variable elements; (*b*) a list of replacement sets for the variable elements; and (*c*) a set of substitution rules that specified which replacement sets could fill which variable element positions. Problems were generated by assigning replacement sets to their corresponding variable element positions, while the fixed element positions remained constant. The item forms, replacement sets, and substitution rules completely defined the universe from which the experimental problems were drawn.

The item forms were based on a subset of the general equations defined by Suppes et al. (1967, p. 168). These number problem equations are shown in the leftmost column of Table 1. The number equations differ with respect to the position of the unknown set (starting or ending) and the sign of the operation (plus or minus). Each number equation appears twice, once for word problems in the forward order of mention (Forms

TABLE 1

NUMBER PROBLEM ITEM FORMS, WORD PROBLEM ITEM FORMS, AND SAMPLE PROBLEMS

Order of mention		Number problem item forms	Word problem item forms	Sample problems
Forward		1. s + c = __	If *subject* started out with *s objects* and he *gained c objects*, how many *objects* did he end up with?	If *Paul* started out with *5 boats* and he *bought 3 boats*, how many *boats* did he end up with?
		2. s − c = __	If *subject* started out with *s objects* and he *lost c objects*, how many *objects* did he end up with?	If *John* started out with *7 cars* and he *sold 2 cars*, how many *cars* did he end up with?
		3. __ + c = e	How many *objects* did *subject* start out with if he *gained c objects* and he ended up with *e objects*?	How many *stamps* did *Dave* start out with if he *found 2 stamps* and he ended up with *6 stamps*?
		4. __ − c = e	How many *objects* did *subject* start out with if he *lost c objects* and he ended up with *e objects*?	How many *balls* did *Bill* start out with if he *lost 2 balls* and he ended up with *5 balls*?
Backward		5. s + c = __	How many *objects* did *subject* end up with if he *gained c objects* and he started out with *s objects*?	How many *boats* did *Paul* end up with if he *bought 3 boats* and he started out with *5 boats*?
		6. s − c = __	How many *objects* did *subject* end up with if he *lost c objects* and he started out with *s objects*?	How many *cars* did *John* end up with if he *sold 2 cars* and he started out with *7 cars*?
		7. __ + c = e	If *subject* ended up with *e objects* and he *gained c objects*, how many *objects* did he start out with?	If *Dave* ended up with *6 stamps* and he *found 2 stamps*, how may *stamps* did he start out with?
		8. __ − c = e	If *subject* ended up with *e objects* and he *lost c objects*, how many *objects* did he start out with?	If *Bill* ended up with *5 balls* and he *lost 2 balls*, how many *balls* did he start out with?

1–4) and once for word problems in the backward order of mention (Forms 5–8).

The item forms for word problems are in the center column of Table 1. Item forms 1–4 present information about the sets in the forward order of mention, that is, in chronological order. Item forms 5–8 present the information in the backward order of mention, that is, in reverse chronological order. The ending set is unknown for Item Forms 1, 2, 5, and 6, and the starting set is unknown for Forms 3, 4, 7, and 8, in accordance with the number problem upon which the item form was based.

TABLE 2

LISTS OF THE REPLACEMENT SETS FOR THE VARIABLE ELEMENTS OF THE ITEM FORMS

Subject	Gain verb	Loss verb	Direct object	Indirect object	Number pair
Bill	bought	sold	balls	sister	5, 3
Dave	found	lost	blocks	brother	5, 2
Dick	took	gave	boats	father	6, 2
Fred	got	sent	books	mother	7, 2
John			cars		
Mark			pens		
Pete			stamps		
Paul			trucks		

Similarly, Forms 1, 3, 5, and 7 accept gain verbs, and Forms 2, 4, 6, and 8 accept loss verbs, in accordance with the sign of the operation of their underlying number problem. Restrictions were applied such that the verbs *bought, found, sold,* and *lost* always appeared in sentences with no indirect object, while *took, got, gave,* and *sent* always appeared in sentences containing an indirect object phrase. Sentences with and without indirect object phrases were equally distributed across all experimental conditions.

The replacement sets used to generate item sentences are shown in Table 2, and a sample problem for each item form is shown in the rightmost column of Table 1. The variable elements of the item forms and the replacement sets substituted in the sample items are shown in italics for the sake of clarity. During item generation, number pairs were selected so that the change set was always smaller than the other given set, that is, the starting or the ending set. This made it possible to either add or subtract the change set, and thus prevented the subjects from obtaining the correct solution operation by simply eliminating one of the two possible operations. The values for s, c, and e ranged from 2 to 9, as suggested by the number pairs shown in Table 2. The gain and loss verbs were chosen to be "minimal contrast" pairs, that is, *bought* and *sold, found* and *lost, took*

and *gave*, and *got* and *sent*[a]. Theoretically, this selection procedure limited the difference between gain and loss verb item forms to the single semantic feature, "acquisitiveness" (cf. McNeill, 1966; Perfetti, 1968). The aim was to make problems taking gain and loss verbs as semantically similar as possible.

A total of 32 word problems were generated, four from each item form. Replacement items were assigned to the variable elements of the item forms so that (*a*) the four number pairs were fully, and therefore, equally represented in the four problems generated from each item form; and (*b*) the four gain and the four loss verbs were fully, and therefore, equally represented in the four problems generated from each gain and loss item form. The experimental problems were typed in large-sized primary characters on plain 5 × 8 inch index cards. The cards were arranged into four blocks of eight problems. Each block contained one problem from each item form. Problems within a block were arranged randomly, and problem blocks were presented in random order.

Apparatus

The apparatus consisted of a plexiglass response box and a 1/100 second timer. The side of the box facing the child sloped upwards at an angle of 60 degrees, and a rack on the upper portion held the current problem card. The lower portion contained a set of 10 push-button microswitches arranged in an arc. The microswitch caps were numbered from 1 to 10, and were one-half inch in diameter. One of 10 correspondingly numbered indicator lights mounted on the back of the box lighted when a response button was depressed. The lights enabled the experimenter to directly read the subjects' responses. The timer was activated automatically by seating a problem card in the rack, and halted automatically when a response button was depressed. Latencies were recorded from the timer by the experimenter.

Procedure

Prior to the experimental task each subject solved 20 arithmetic number problems arranged in five randomized blocks. Sixteen of these problems were constructed from the number pairs in Table 2, with the unknown position designated by an underline. Four additional problems were constructed, one from each form in Table 1, so that all of the response keys would be included in the set of correct responses. Subjects were instructed to solve a problem and then to press the appropriate answer key with the index finger of the preferred hand. The subjects were told that the object of the task was to get as many problems correct as they could, and that they would be informed of their score after every fourth problem and again at the end of the task. A buzzer pressed by the experimenter served as a ready signal prior to presenting a problem. The warm-up task allowed subjects to become familiar with the apparatus and afforded the opportunity to assess basic computational skills, which were defined as performance on s + c = __ and s − c = __ problems. The subjects in Experiment 1 had an average of 92% correct; the subjects in Experiment 2 had an average of 99% correct.

Immediately following the warm-up task, each subject solved 32 arithmetic word problems. The subjects were instructed to read each problem aloud and, after solving it, to press the appropriate answer key with the index finger of the preferred hand. Subjects were told again that the object of the task was to get as many problems correct as they could and that they would be informed of their score after every eighth problem and again at the end of the task. Subjects were given a three-minute rest when half of the problems were finished. The average time to complete the warm-up and experimental tasks combined was about 25 minutes.

[a] The minimal contrast for the word *sent* should more properly be *receive*, but this word was considered too difficult for third-grade subjects.

Results

Error Data

A 2 × 2 × 2 analysis of variance with repeated measures on every factor was performed for both experiments to determine the effects of identity of the unknown set (starting versus ending), type of verb (gain versus loss), and order of mention (forward versus backward).

The results were almost identical for the two replications. Except for the effect of type of verb which was borderline in Experiment 1, the three main effects were significant in both experiments as follows: identity of the unknown set ($F = 103.09$, $df = 1/28$, $p < .001$; $F = 63.42$, $df = 1/33$, $p < .001$); type of verb ($F = 3.87$, $df = 1/28$, $p < .06$; $F = 9.01$, $df = 1/33$, $p < .01$); and order of mention ($F = 20.22$, $df = 1/28$, $p < .001$; $F = 24.93$, $df = 1/33$, $p < .001$). There were two significant interaction effects in both experiments, identity of the unknown set with order of mention ($F = 18.84$, $df = 1/28$, $p < .001$; $F = 4.73$, $df = 1/33$, $p < .05$), and identity of the unknown set with type of verb ($F = 19.51$, $df = 1/28$, $p < .001$; $F = 8.13$, $df = 1/33$, $p < .01$). Mean number of errors for the cells involved in these interactions are shown in Table 3.

The effect for identity of the unknown set

TABLE 3
MEAN ERRORS AND STANDARD DEVIATIONS FOR SIGNIFICANT INTERACTIONS

Set	Experiment 1		Experiment 2		Experiment 1		Experiment 2	
	Forward order	Backward order	Forward order	Backward order	Gain	Loss	Gain	Loss
Starting unknown								
$\bar{X}$	4.28	6.07	3.47	4.88	6.10	4.24	5.03	3.32
SD	2.28	2.09	2.29	2.37	2.09	2.43	2.53	2.67
Ending unknown								
$\bar{X}$	1.38	1.41	.79	1.38	.93	1.86	1.03	1.15
SD	1.42	1.38	.91	1.26	1.22	1.98	1.27	1.18

was as predicted (Hypothesis 3): there were more errors when the unknown set was the starting set ($\bar{X}$ = 10.34 versus 2.79 in Experiment 1; 8.35 versus 2.18 in Experiment 2). In accord with Hypothesis 1, there were more errors when the order of mention was backwards ($\bar{X}$ = 7.48 versus 5.66 in Experiment 1; 6.26 versus 4.26 in Experiment 2). Gain verb problems were also more difficult than loss verb problems ($\bar{X}$ = 7.03 versus 6.10 in Experiment 1; 6.06 versus 4.47 in Experiment 2). The latter two effects, however, were due mainly to the greater difficulty of backward order and gain verb problems when the starting set was the unknown set, as is shown in Table 3.

Reaction Time

Response latencies were used as a basis for inferring the processes employed for correctly solving the different classes of problems. The reaction time data is based on correctly solved problems in the second through the fourth blocks; the first block was eliminated to achieve more stable reaction times. Two latencies in Experiment 1 exceeded 30 seconds (59.92 and 57.55 seconds). They were classified as outliers and excluded from the data analysis on the assumption that a different process was used in solving these problems than others of the same class. The means for each subject in each cell of the design were calculated for the remaining latencies. Due to the high error rate in some cells, only four of the 29 subjects in Experiment 1 and 12 of the 34 subjects in Experiment 2 had a full complement of cell means and could be included in the data analysis.

A 2 × 2 × 2 analysis of variance with repeated measures on every factor was performed on the latency data for both experiments. Only the identity of the unknown set emerged as a clearly significant main effect in both experiments (F = 12.82, df = 1/3, $p < .05$; F = 21.93, df = 1/11, $p < .001$). The mean reaction times for starting set and ending set unknown problems were 17.18 versus 14.43 seconds in Experiment 1; 13.88 versus 11.68 seconds in Experiment 2. Thus, Hypothesis 4 was supported. However, Hypothesis 2 was not, since there was no significant effect for order of mention. A statistically weak (F = 2.99, df = 1/11, $p < .12$) effect of order of mention occurred in Experiment 2, with backward order problems, taking approximately one second longer to solve than forward order problems ($\bar{X}$ = 13.26 versus 12.30 seconds). The interaction between the Identity of the Unknown Set × Type of Verb approached significance in Experiment 2 (F = 4.42, df = 1/11, $p < .06$), and is plotted in Figure 1. As the graph makes clear, gain verb problems took substantially longer when the unknown set was the starting set, while they took somewhat less time when the ending set was unknown.

DISCUSSION

Of the four hypotheses initially stated, three were clearly supported. Backward order of mention problems produced more errors than did forward order problems (Hypothesis 1). Problems with an unknown

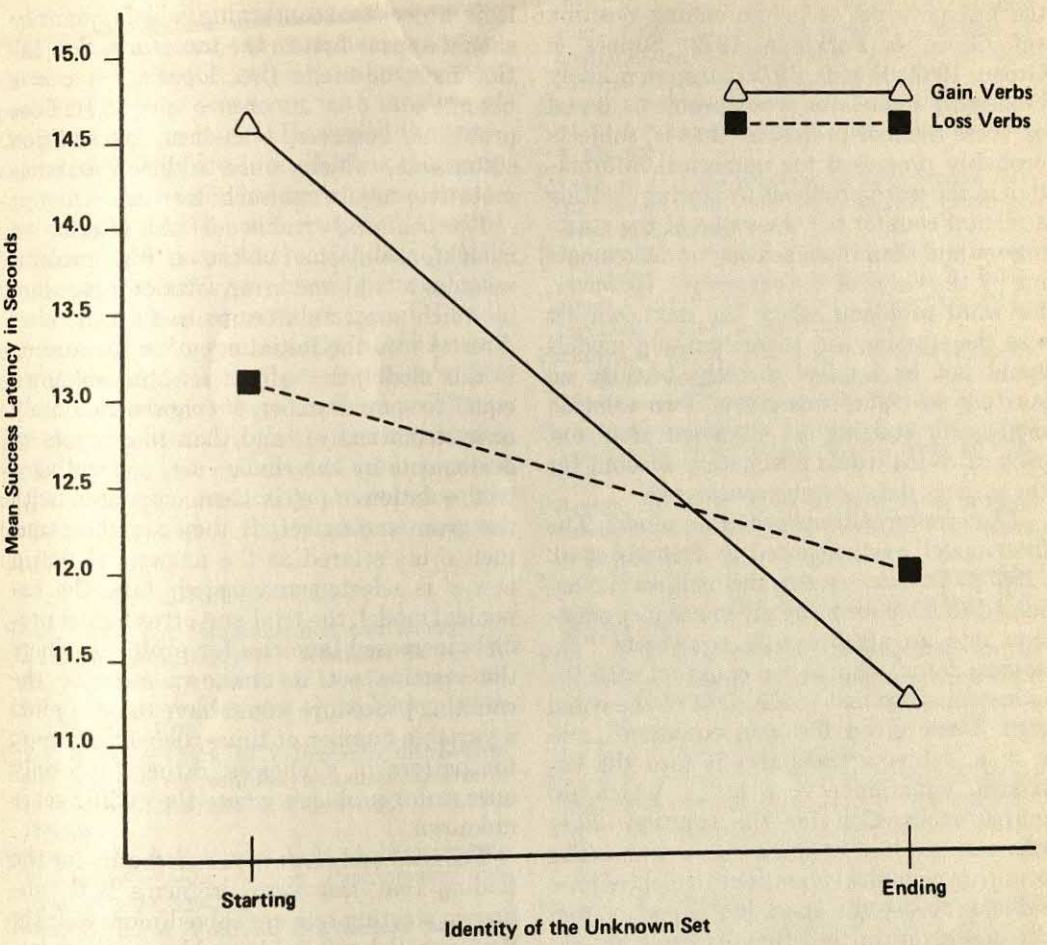


FIGURE 1. Reaction time data for Experiment 2: Interaction of identity of the unknown set with type of verb for ($N = 12$).

starting set produced both more errors (Hypothesis 3) and longer latencies (Hypothesis 4) than did problems with the ending set unknown. There was no clearly significant latency effect for the order of mention variable (Hypothesis 2).

Order of Mention Variable

The main effect of the order of mention on the error data was accompanied by an interaction with the identity of the unknown set (see Table 3). This interaction is difficult to interpret since it reflects a virtual absence of errors for ending set unknown problems. These problems may have been so easy as to overpower the effect of order of mention. In such a case, there would really be only a main effect of order of mention to explain. To find out whether there is a *true* interaction of these variables, what is needed is further experimentation in which both error and latency effects can be examined under conditions in which ending set unknown problems are not so easy to solve.

Identity of the Unknown Set Variable

The identity of the unknown set emerged as the most powerful variable affecting problem solution in both experiments. When the unknown set was the starting set, problems were more difficult to solve. There were more errors, and successful solutions took longer. The counting models that describe processes for solving number problems where

the unknown set is in the ending position (cf. Groen & Parkman, 1972; Suppes & Groen, 1967; Woods, 1972) suggest a likely processing model for word problems based on these number problems; that is, subjects probably processed the numerical information in the word problems by storing (setting a mental counter to) the value of the starting set and then incrementing or decrementing by the value of the change set. However, for word problems where the unknown set was the starting set, these counting models could not be applied directly because no starting set value was given. Two solution models for starting set unknown problems, each of which would adequately account for the present data, can be considered.

The canonical transformation model. The first model was suggested by Suppes et al. (1967). In this model the subject is assumed to transform the given number equation into an algebraically equivalent "canonical form", that is, an equation with the unknown alone and to the right of the equal sign. Thus, given the gain equation $__ + c = e$, subjects transform it into the canonical equation $e - c = __$, which requires subtraction for the solution; likewise $__ - c = e$ becomes $e + c = __$. The resulting canonical equations, which require subjects to use the operation opposite from the one given in the stimulus problem, are then solved using a counting model. The canonical model predicts more errors and longer solution times for problems where the unknown set is the starting set because an additional process for transforming the given equation is required.

Suppes et al. (1967) found that starting set unknown equations with subtraction signs (i.e., $__ - c = e$) took less time to solve and caused fewer errors than those with addition signs. Our word problem data parallel these findings. For starting set unknown problems, gain verbs caused more errors than loss verbs (Table 3), and took longer to solve (Figure 1). The greater difficulty of gain over loss problems can be accounted for within the canonical model by assuming that the transformation is more complex for gain verb problems. This assumption is reasonable, since for gain problems a decision concerning which quantity should appear first in the transformed equation must be made (i.e., Does $__ + c = e$ become $e - c = __$, or $c - e = __$?). Loss problems, however, transform to addition statements, which, since addition is commutative, require no such decision.

The trial and error model. An alternative model for solution of unknown first problems assumes a trial and error solution procedure in which subjects attempt to fit a number directly into the initial equation. According to this model, the subject sets the unknown equal to some number, s′ (chosen randomly or systematically), and then increments or decrements by the change set, c, yielding a trial solution e′. e′ is then compared with the given ending set; if they are the same, then s′ is declared as the answer; if not, a new s′ is selected and tested. Like the canonical model, the trial and error model predicts increased latencies for problems where the starting set is unknown because the counting procedure would have to be applied a variable number of times, depending upon the pattern of s′ choices, rather than only once as for problems where the ending set is unknown.

The trial and error model accounts for the finding that loss verb problems with unknown starting sets are solved more quickly than parallel gain verb problems in the following way: Loss verb problems yield a given equation of the type $__ - c = e$; this form of equation permits subjects to make the inference that the value of the starting set, s′, must be equal to or greater than the change set, c. By contrast, in equations of the type $__ + c = e$, the change set provides no constraint on the possible values of s′. For loss verb problems, the constraint on the possible values of s′ may facilitate the selection of an initial value with which the trial and error algorithm can be initiated and may thus allow the trial and error process to begin more quickly. Also, this constraint limits the area of search for the unknown value, so that fewer iterations of the trial and error process are needed. For gain verb problems, the subject must estimate a good starting value for s′ and will also, on the average, test more values.

As the above discussion suggests, both the trial and error and the canonical transformation models for solving unknown first problems appear consonant with the data of the present experiment. However, the present latency data are limited, thus requiring summation across individuals whose strategies may in fact be different, and also introducing considerable variability which limits the possibility of drawing strong conclusions with regard to either model. In addition, it is unknown exactly how the linguistic demands of the present task might affect processing strategy for these problems. For these reasons further experiments designed to test the validity of both models, perhaps using numerically presented equations rather than verbal problems, are required.

REFERENCES

Bever, T. G. The comprehension and memory of sentences with temporal relations. In G. B. Flores d'Arcais & W. J. M. Levelt (Eds.), *Advances in psycholinguistics*. New York: American Elsevier, 1970.

Clark, H. H., & Clark, E. V. Semantic distinctions and memory for complex sentences. *Quarterly Journal of Experimental Psychology*, 1968, **20**, 129–138.

Groen, G. J., & Parkman, J. M. A chronometric analysis of simple addition. *Psychological Review*, 1972, **79**, 329–343.

Groen, G. J., & Poll, M. Subtraction and the solution of open sentence problems. *Journal of Experimental Child Psychology*, 1973, **16**, 292–302.

Hively, W. II, Patterson, H. L., & Page, S. H. A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, **5**, 275–290.

Huttenlocher, J., & Weiner, S. L. Comprehension of instructions in varying contexts. *Cognitive Psychology* 1971, **2**, 369–385.

Loftus, E. F., & Suppes, P. Structural variables that determine problem-solving difficulty in computer-assisted instruction. *Journal of Educational Psychology*, 1972, **63**, 531–542.

McNeill, D. A study of word association. *Journal of Verbal Learning and Verbal Behavior*, 1966, **5**, 548–557.

Osburn, H. G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, **28**, 95–104.

Paige, J. M., & Simon, H. A. Cognitive processes in solving algebra word problems. In B. Kleinmuntz (Ed.), *Problem solving: Research, method, and theory*, New York: Wiley, 1966.

Parkman, J. M., & Groen, G. J. Temporal aspects of simple addition and comparison. *Journal of Experimental Psychology*, 1971, **89**, 335–342.

Perfetti, C. A. Minimal contrast in artificial word associations. *Psychonomic Science*, 1968, **10**, 229–230.

Peterson, M. J., & Aller, S. Arithmetic problem solving. *Journal of Experimental Psychology*, 1971, **91**, 93–97.

Restle, F. Speed of adding and comparing numbers. *Journal of Experimental Psychology*, 1970, **83**, 274–278.

Steffe, L. P. The effects of two variables on the problem-solving abilities of first-grade children. (Tech. Rep. 21) Madison: University of Wisconsin, Wisconsin Research and Development Center for Cognitive Learning, 1967.

Suppes, P., & Groen, G. Some counting models for first grade performance on simple addition facts. In J. M. Scandura (Ed.), *Research in mathematics education*. Washington, D. C.: Council of Teachers of Mathematics, 1967.

Suppes, P., Hyman, L., & Jerman, M. Linear structural models for response latency performance in arithmetic on computer-controlled terminals. In J. P. Hill (Ed.), *Minnesota symposia on child psychology*, Vol. 1. Minneapolis: University of Minnesota Press, 1967.

Woods, S. An experimental test of five process models for subtraction. Unpublished master's thesis, University of Pittsburgh, 1972.

(Received September 25, 1973;
Revised April 9, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 826–834

RELATIONSHIP BETWEEN MENTAL ABILITIES, SOCIAL CLASS, AND EXPOSURE TO ENGLISH IN CHINESE FOURTH GRADERS[1]

LELAND Y. YEE[2] AND ROLFE LAFORGE

California State University, San Francisco

This study attempted to examine the relation between mental abilities as defined by *raw scores* on the Wechsler Intelligence Scale for Children (WISC), social class, and exposure to English in 53 American-born Chinese 9- to 10-year-olds attending a private school in San Francisco. From canonical, stepwise, and multiple regression analyses, no relation was found between social class and the 12 WISC subtests, the WISC Overall total, the six WISC Verbal subtests, the WISC Verbal total, or the WISC Performance total. A significant relation was found, however, between social class and the six WISC Performance subtests. There was also a small but significant relation between the exposure to English variables and the WISC Overall total. Among the possible explanations proposed for the unexpected results were: small, possibly biased sample, inappropriate social class measures, and the nature of social class and its relation to WISC scores within the San Francisco Chinese-American population.

Definitive statements on social class differences in intelligence are sparse. Hess (1970) attributed this to the paucity of adequate theories to guide such research and to the following methodological problems: the use of different socioeconomic status (SES) measures, the absence of definitive social class categories, the absence of attempts to partial out the effects of variables related to social class, and the lack of research on intrasocial class performance variability.

Hess (1970) claimed that there is a disparity range of 8 to 23 IQ points between upper and lower social class groups, depending on the test used and the subject's age level. Curry (1962), in an attempt to identify the source of SES differences on intellectual performance, divided his sixth-grade subjects into different SES (lower, middle, and upper) and intelligence levels (low, medium, and high—California Test of Mental Maturity). He found several SES differences in reading and language achievement. Where they were significant, the higher SES groups consistently had the higher scores. The only exception was with the "upper intelligence group." Curry suggested at that intelligence level, the effects of SES differences are masked by the high intelligence.

Lesser, Fifer, & Clark (1965), using a revised version of the Hunter College Aptitude Scales for Gifted Children, found that middle-class Chinese first graders (six to seven years old) scored significantly higher on the Verbal, Reasoning, Number, and Space scales than did lower-class Chinese first graders. However, the patterns of these scores remained the same for the two groups. To determine whether such differences between SES groups existed in even older subjects, Backman (1971), using twelfth-grade subjects from Project Talent, a nationwide study conducted in 1960, found significant differences in both the patterns and levels of mental abilities between lower-middle

[1] This study was revised for publication from a thesis completed by the senior author at California State University, San Francisco.

The authors wish to thank Harry Osser and Becky Loewy for their assistance in this study.

[2] Requests for reprints should be sent to Leland Y. Yee, Department of Psychology, University of Hawaii, Honolulu, Hawaii 96822.

and upper-middle SES subjects. However, Backman (1971) noted that although there was a significant difference on the patterns of mental abilities across SES groups, the difference was too small to be considered important.

It should be noted that most of the relatively few studies on the relation between social class and intelligence have relied on black subjects and/or white subjects (Christiansen & Livermore, 1970), with very few using Chinese subjects (Hess, 1970). An exception was the Lesser et al. (1965) study which found that different ethnic groups (Jewish, Puerto Rican, Chinese, and Negro) exhibited different levels and patterns of mental abilities. Other experimenters have also found these group differences (Anastasi, 1967; Eells, Davis, Havighurst, Herrick, & Tyler, 1951; Hess, 1970).

The literature on whether bilingualism is an aid or a hindrance to intelligence test performance (Peal & Lambert, 1962; Soffietti, 1955) is also equivocal. Darcy (1963), in her review of the effects of bilingualism on the measurement of intelligence, summarized the reasons for the inconsistent findings as follows: (*a*) Subjects are introduced to the second language at disparate ages. (*b*) Subjects' SES and cultural backgrounds are not matched. (*c*) Measurements of degrees of bilingualism are not uniform. (*d*) Intelligence tests used vary. (*e*) Different methods are used to teach subjects the second language. (*f*) Differential language handicap is confounded with educational retardation.

It was hypothesized that: (*a*) there is a relation between social class and mental ability as measured by the 12 WISC subtests (Lesser et al., 1965, p. 52); (*b*) there is a relation between social class and the six Verbal subtests, but no demonstrable relation to the six Performance subtests (Lesser et al., 1965, p. 5); (*c*) there is a positive multiple correlation between social class and the Overall total, and between social class and the Verbal total, but no demonstrable multiple correlation between social class and the Performance total (Lesser et al., 1965, p. 5); (*d*) exposure to English contributes importantly to the multiple correlation between social class and the Overall total, and to the multiple correlation between social class and the Verbal total (Lesser et al., 1965, p. 9).

A multivariate analysis was used to test these hypotheses. For the first two hypotheses, relating social class and exposure to English to WISC scores, the dependent variables were the 12 subtests. These were separated into two subsets: six Verbal and six Performance subtests. For the last two hypotheses, relating social class and exposure to English to WISC totals, the dependent variables were the Overall, Verbal, and Performance totals.

Method

Subjects

The sample consisted of 53 American-born Chinese fourth graders (29 girls and 24 boys) from one of the Freedom Schools[3] in Chinatown, San Francisco. Their mean age was 9½ years.

Variables

Social class. A social class index was used in preference to an SES index; strictly economic measures were omitted after considering two earlier studies. In one study, Kahl & Davis (1965) factor analyzed 19 stratification indices and found that parent's occupational level was the dominant variable, with education and residence (home types) as the next most important variables. Kohn (1969) also concluded that income adds very little to identifying class position. Since the Chinese people seldom divulge their income to anyone, a social class index also avoided the near impossibility of ascertaining the income of the subjects' families.

The basis for designating social class came from three variables: Residence, education, and occupation of the father or head of household. Residence was categorized on a 6-point scale (6 = ownership; 5–1 = decreasing amounts of rent: \$170 or more, \$150–\$169, . . ., \$109 or less). Education was categorized in two parallel 8-point scales in an attempt to equate education in Hong Kong with education in the United States. The occupation category was a 6-point scale derived mainly from Lesser et al. (1965). Occupations not in the list of Lesser et al. (1965) were classified according to

[3] Late in 1971, in response to feelings against busing, Chinese parents banned together to form several private schools which have come to be known as Freedom Schools. Subjects were recruited from a Freedom School, rather than a public school in San Francisco, because of a complete ban on any type of intelligence testing in San Francisco public schools.

their similarity (as defined by the U.S. Government's *Dictionary of Occupational Titles* (1965) with jobs found in the list. Thus, a job title not found in the list of Lesser et al. (1965) was given the same category number as other job titles appearing with it in the *Dictionary of Occupational Titles* (1965). Thus, although the social class indices were traditional indices used by other investigators, the content represented the unique social character of the Chinese in Chinatown, San Francisco.

Exposure to English. Thirty variables related directly or indirectly to exposure to English were assessed by means of a parent interview. From these, four variables were selected, mainly by factor analysis, to represent the exposure to English domain (see *Results* section).

Procedure

The race of the tester and testee was matched in this study, as both were Chinese.

A month was spent becoming acquainted and establishing rapport with all of the fourth graders in the Freedom School. During recess and physical education periods, the experimenter organized relays and other group games. In the classroom, he helped each student with his classwork.

Following that month, the teachers of these students were asked, "Can (the student) carry on a normal conversation in English?" and "Do you consider his understanding of the English language adequate for his age?" Also, students were asked their birthplace. If the first two questions were answered affirmatively and if the student was U.S. born, he was later given the WISC (in English). These three criteria were used to provide some assurance that any subject's response would not be a function of his lack of understanding the English language. Birthplace was confirmed during the parent interview. From a total of 84 fourth graders, 53 satisfied the criteria.

Following the administration and scoring of the WISC, each parent was interviewed. The interview schedule consisted of five sections (family structure, school experience, bilingualism, how subjects spent their free time during weekdays and weekends, and information relevant to the education, occupation, and residence categories). A conversation with the parents prior to administering the interview schedule was held to decide whether to use the English or the Chinese version.

Results

All families were intact, except for three where the father was deceased. The subjects had, on the average, .94 brothers, 1.2 sisters, and .21 older relatives (typically grandparents) living with their family. Forty-one subjects had attended one of the two public elementary schools at the center of Chinatown (where over 90% of the students are Orientals) prior to their present enrollment in the Freedom School. On the average, the subjects had attended 3.5 months of nursery school and 2.2 years of Chinese language school. (These schools are in session daily after English school.) The average percentage of English spoken at home varied from 75.6% (subject) to 28.3% (father) to 16.8% (mother). All the parents believed that bilingualism was important, but 34 felt that English should be given greater emphasis than Chinese.

Parents reported an average of 4.8 hours of free time available for the subjects on a weekday and 11.6 hours during the weekend. Of those amounts, 2.9 hours were spent watching television during a weekday and 5.1 hours during the weekend. The subjects went approximately 1.3, .17, and 2.3 times per month to a Chinese movie, to an American movie, and to the library, respectively. Eight other exposure to English variables such as amount of free time spent reading English or Chinese books, listening to the radio, etc. provided little response variability and were dropped from further analysis. On the social class variables, it was found that the average category was 2.2 for occupation (e.g., skilled worker), 3.0 for parent education (eleventh grade), and 3.5 for residence ($150 rent).

Three exposure to English variables: (*a*) percentage of English spoken at home by the subject's older brothers and sisters, (*b*) percentage of English spoken at home by subject's younger brothers and sisters, and (*c*) percentage of English spoken at home by subject's older relatives, were not considered in the analysis because not all subjects had brothers, sisters, or older relatives. However, the influence of the languages used by the siblings and older relatives was represented by the variable, *average percentage of English spoken at home by the parents, siblings, and older relatives.*

Several factor analyses were performed to reduce the number of exposure to English and social class variables before correlating them with the WISC scores. The factor analysis which combined the remaining 19 exposure to English variables with the three social class variables is reported. All prin-

TABLE 1

FACTOR LOADINGS OF THE EXPOSURE TO ENGLISH AND SOCIAL CLASS VARIABLES

Variables	Factors							
	1	2	3	4	5	6	7	8
Sex of subject (0 = female, 1 = male)	.14	−.15	−.*52*	−.*50*	−.07	−.27	−.26	.19
Presence of father	.11	−.11	−.01	−.03	.*77*	.07	.10	−.05
No. brothers	.29	.08	.07	−.19	.09	−.02	−.06	−.*72*
No. sisters	.04	−.00	.27	−.10	−.03	.*83*	.07	.14
Percentage of English-speaking students in subject's school	.01	.11	.*85*	−.09	−.01	.08	−.07	−.24
Six months period in nursery school	.03	−.08	−.10	.*83*	.03	−.14	−.10	.16
Years in Chinese school	−.44	−.02	−.07	.*51*	.02	.10	.*51*	.12
Percentage of English spoken at home by subject	.*62*	.00	−.25	.17	−.30	−.02	−.27	−.33
Percentage of English spoken at home by father	.*86*	−.07	.04	.09	.16	.08	.05	−.09
Percentage of English spoken at home by mother	.*75*	.04	−.11	−.17	.16	−.11	−.03	.06
Average percentage of English spoken at home by subject's parents, siblings, and older relatives	.*91*	.02	.04	−.08	−.18	.15	−.10	−.17
Parent's emphasis of either Chinese (0) or English (2) or both (1)	−.14	−.13	.24	−.01	.03	−.11	.07	−.*75*
Hours of free time subject has during a weekday	.12	−.*50*	.10	−.01	.02	.02	−.*65*	−.07
Hours of free time subject has during a weekend	.00	−.10	−.03	.09	−.05	−.04	−.*86*	.08
Hours subject spent watching television during a weekday	.14	−.*68*	−.14	−.17	.13	−.12	−.35	−.21
Hours subject spent watching television during a weekend	−.10	−.*82*	−.10	.10	−.02	.02	−.05	.02
No. times in a month subject goes to Chinese movies	−.12	−.08	−.*46*	.02	.21	.*65*	.09	.01
No. times in a month subject goes to American movies	.06	−.33	.10	−.20	−.*50*	−.32	.26	.20
No. times in a month subject goes to the library	−.01	.32	.17	.*46*	−.06	.*48*	−.30	.08
Occupation of head of household	.*63*	.15	.28	−.18	.34	−.22	−.04	.18
Education of head of household	.*59*	.08	.03	.13	.*48*	−.20	−.02	.19
Residence of head of household	.27	.*46*	−.01	−.02	.42	−.11	.07	−.25

Note. Italics indicate factor loadings of greater than .45. Factor 1 = Average percentage of English Spoken at Home by Subject's Parents, Siblings, and Older Relatives; Factor 2 = Free Time, Including Time Spent Watching TV; Factor 3 = Distance from Chinatown; Factor 4 = Formal Instruction Outside Elementary School.

ciple components with latent roots greater than one were extracted and rotated according to the varimax criteria (Morrison, 1967, pp. 229, 286). The latent roots greater than one were 4.26, 2.73, 1.92, 1.88, 1.45, 1.34, 1.14, and 1.07. The cumulative percentage of variance accounted by these eight principal components was 72%. The rotated factors appear in Table 1.

Exposure to English variables with the highest loading on each rotated factor were selected as a "cluster" to represent that factor (see variables in italics in Table 1). It was decided that the 19 exposure to English variables would be represented by four clusters: Average Percentage of English Spoken at Home by Subject's Parents, Siblings, and Older Relatives (AV %); Free Time, Including Time Spent Watching TV (FT); Distance from Chinatown (D-CHINA); and Formal Instruction Outside Elementary School (FOR-INST). Only four clusters were selected because factors beyond the fourth were either uninterpretable (Fac-

tor 6), overlapping with earlier factors (Factor 7), or representative of very few subjects (Factors 5 and 8). Each cluster score was the simple sum of the exposure to English variables loading above .45 on the corresponding varimax rotated factor.

In the social class domain, occupation and education correlated .55; they correlated .42 and .33, respectively, with residence. Hence, occupation and education (occupation–education) were added together to form a single summary social class variable, while residence was retained as a second social class variable. These two social class variables and the four exposures to English clusters were used thereafter in the analyses as independent variables to relate to the dependent variables (12 subtests, six Verbal subtests, six Performance subtests, and three totals).

The intercorrelations of the exposure to English and social class variables were generally low. The six intercorrelations of the four exposure to English variables were less than or equal to .15, except for FOR-INST with AV % (−.25) and with D-CHINA (−.17). Five of the nine correlations involving the social class variables were less than .15; only two were at all large, both involving occupation–education (.44 with AV % and .43 with residence). Residence itself correlated .19 with AV % and −.22 with FT.

The intercorrelations of the six Verbal subtests ranged from zero (.01) to a moderate high (.62), with Vocabulary showing the highest correlation to the other Verbal subtests (.41 to .62) except with Digit Span (−.13). Digit Span also showed essentially zero correlations with all other subtests. The intercorrelations of the Performance subtests were less than .30 in all but one case; Block Design correlated .40 with Object Assembly. Coding and Mazes showed essentially zero correlations with all other subtests. The correlations of Verbal with Performance subtests were likewise low, with only six in the .30 to .42 range, while Similarities and Block Design showed the highest correlations.

The correlations between the six independent variables and the WISC subtests were very low, with the highest between D-CHINA and Arithmetic at −.42; with Information at −.31; and Vocabulary at −.31. A correlation of .32 was found between FOR-INST and Picture Arrangement, and FT with Comprehension, .34, and .36 with Mazes.

To test the significance of the relation between the six independent variables and the various subsets of the 12 WISC variables,

TABLE 2

INTERCORRELATIONS OF THE COMBINED EXPOSURE TO ENGLISH, COMBINED SOCIAL CLASS, AND WECHSLER INTELLIGENCE SCALE FOR CHILDREN (WISC) TOTALS

Predictors	Exposure to English				Social class		WISC		
	1	2	3	4	5	6	7	8	9
Exposure to English									
1. AV %	1.00	.13	.15	−.25	.44	.19	.05	−.16	−.10
2. FT	.13	1.00	−.04	−.10	−.01	−.22	.26	.14	.22
3. D-CHINA	.15	−.04	1.00	−.17	.09	.05	−.35	−.35	−.42
4. FOR-INST	−.25	−.10	−.17	1.00	−.13	−.06	.15	.33	.32
Social Class									
5. Occupation–Education	.44	−.01	.09	−.13	1.00	.43	.06	−.03	.00
6. Residence	.19	−.22	.05	−.06	.43	1.00	.14	−.16	−.06
WISC									
7. Verbal Subtest Total	.05	.62	−.35	.15	.06	.14	1.00	.33	.71
8. Performance Subtest Total	−.16	.14	−.35	.33	−.03	−.16	.33	1.00	.90
9. Overall Total	−.10	.22	−.42	.32	.00	−.06	.71	.90	1.00

Note. AV% = Average Percentage of English Spoken at Home by the Subject's Parents, Siblings, and Older Relatives; FT = Free Time, Including Time Spent Watching TV; D-CHINA = Distance from Chinatown; FOR-INST = Formal Instruction Outside Elemetary School.

the procedure outlined by Morrison (1967, p. 212) for testing the independence of K sets of variables was used. To answer the first hypothesis entailed the computing of three determinants: the intercorrelations of the 12 dependent variables, the six independent variables, and all 18 variables. The result was nonsignificant ($\chi^2 = 76.32$, $df = 72$, $p < .05$); no relation was demonstrated between the 12 WISC variables and the six independent variables. A similar test of the first part of the second hypothesis produced a similar result; no relation was demonstrated between the six independent variables and the six Verbal subtests ($\chi^2 = 49.76$, $df = 36$, $p < .05$). The latter part of the second hypothesis was rejected ($\chi^2 = 78.72$, $df = 36$, $p < .05$). Thus the predictions made in both parts of the second hypothesis were reversed by statistical findings.

To examine the third hypothesis, three separate stepwise regression analyses were performed. All F values were nonsignificant. The six correlations between two social class and the three WISC totals were less than .17. Neither social class measure correlated significantly with the three totals. Therefore, the first two parts of the third hypothesis were not supported, but the third part, that social class would *not* be significantly related to the scores on the Performance subtests, was confirmed.

Finally, to test the fourth hypothesis, three multiple regression analyses were performed. It was found that there was a significant relation ($p < .05$) between the six independent variables and the Verbal total. These same six independent variables also significantly ($p < .05$) predicted the Performance total, and the Overall total ($p < .01$). All three F values barely equal the significance value. The fourth hypothesis as stated depended on a relation between social class and WISC; but no such relation appeared. Moreover, the three correlations of AV % with the WISC totals were all less than .17. At most, it can be said that there was a relation between the other three exposure to English variables and the WISC totals. From stepwise regression analysis, one exposure to English variable (D-CHINA) correlated −.35 with the Verbal total. When two variables were entered (FT and D-CHINA), the multiple correlation with Verbal total was .410. D-CHINA correlated −.35 with the Performance total, and −.42 with the Overall total. Including FOR-INST resulted in a multiple correlation of .47 with the Overall total. (All multiple correlation coefficients cited were adjusted for df.)

Discussion

The present study was not expected to reject, or at most, find only borderline acceptance of all hypotheses. Previous studies have demonstrated significant differences in intellectual abilities between individuals from different social class levels (Burnes, 1970; Eells et al., 1951; Lesser et al., 1965).

Several explanations were considered but rejected through further analyses. It was noted earlier that because education and occupation correlated highly (.55), their sum was used as a single social class variable. Consideration was given that this procedure might have contributed to the nonsignificant results. However, for occupation, the highest correlation with a WISC subtest was with Picture Arrangement (.12); the lowest was with Information (−.02). For education, the highest was again with Picture Arrangement (.15); and the lowest was with Arithmetic (.06). Thus, no information was lost in the process of combining the two variables.

Another possible explanation for the low correlations was sample homogeneity. To evaluate this possibility, all the subjects' raw scores were transformed into IQ scores. The mean IQs were 98.6, 110.7, and 104.8 for Verbal, Performance, and Overall scores, respectively. This is contrasted with the mean IQ of 100 established by Wechsler. The standard deviations were 17.6, 18.7, and 17.5 for the Verbal, Performance, and Overall scores, respectively. These compare closely to the standard deviation Wechsler designated for his test, 15.

With these rejected explanations, there is still a number of remaining explanations. Referring again to homogeneous sample, it is possible, of course, that the sample is *more* homogeneous in social class or exposure

to English than the Chinese-American society as a whole, since all parents in the sample chose to send their children to a Freedom School rather than have them bused to public schools lacking a Chinese-American orientation. In addition, due to the small *N*, many of the factor loadings in the 50s and 60s could be the result of chance fluctuation. To give some impression of the magnitude of uncertainty, consider that a .95 confidence interval for estimating a single correlation coefficient would allow an error of approximately ±.28.

The low correlation between social class and WISC scores could be the product of inappropriate measures to identify different social classes. Possibly, the Residence scale did not adequately separate individuals along the social class dimension. Rents are high in the Chinatown area, but parents with limited English ability continue to live there because of the conveniences and security. Chinatown is a self-contained community. The food stores, department stores, banks, theatres, and parks are within walking distance. Until recently, violence in the streets was extremely rare.

The Occupation scale may also be inappropriate because the status value of each occupation was not derived from the values and opportunities existing for this particular subculture. For example, there were parents in the sample with high school and/or college education who were cooks, while there were other cooks who had only an elementary level education and owned their homes. Similarly, although the Education scale represented an attempt to equate education in the San Francisco area to the education in Hong Kong, the classification criteria were still based on values foreign to the Chinese population.

Finally, the results could be a true reflection of the relation between social class and mental abilities in American-born Chinese, that is, there is little or no relation. Any potential differences might be masked by the parents' attitudes toward learning or compensated by their emphasis on the importance of bilingualism and biculturalism. This idea may also be used to explain the significant relation between the predictor variables (exposure to English and social class) and the Performance subtests. Since most of the parents reported that they spoke very little English at home, they are probably providing little direct opportunity to developing competence in English. However, this lack is compensated by their stress on studying. Valuing study and achievement on tasks set by adults, the child might do well on tasks that are not as closely bound to the English language as are the Verbal subtests. This guess is given more credibility by the .33 correlation between Performance total and the exposure to English variable, FOR-INST.

The only other independent variable that correlated as highly with the WISC scores was the exposure to English variable, D-CHINA, which correlated −.35, −.35, and −.42 with the Verbal total, Performance total, and Overall total, respectively. D-CHINA was the sum of the value given to the number of times per month the subject attended a Chinese movie plus the value given to the public school attended. In computing this cluster score, the public school variable was weighed two because its factor loading was twice the *Chinese movie* variable. Remarkably, the correlation was negative, and thus opposite of what one might expect if exposure to English raised WISC scores. To explain the relation between D-CHINA and the WISC scores, it may be hypothesized that the closer one is to the Chinatown area, the more likely he is to maintain some of the traditional values (to work and to study hard) of the Chinese.

Returning to the FOR-INST variable, one might expect that the more exposure one has to educational institutions, the higher the test scores. However, this held only for the Performance and Overall totals, and not for the Verbal total. The FOR-INST variable consisted of three learning experiences: attendance at Chinese school, nursery school, and going to the library, and they correlated with the Performance total at .42, .28, and −.06, respectively. Thus, it might be that the crucial factor was attendance at Chinese school. If so, this relation is in the opposite

direction to that predicted; emphasis on Chinese, not English, raises Performance total.

The lack of relation between AV % and the Verbal, Performance, and Overall totals is difficult to explain, considering the importance investigators (Bernstein, 1961; Hess & Shipman, 1965) have placed on the verbal exchanges between the parents and the child toward developing his learning skills, that is, what and how he learns. However, if one were to examine some of the English verbal exchanges between the parents and the subjects, one might describe the parents' speech as "restricted" (Bernstein, 1961), that is, English speech characterized by simple sentences with few modifiers and lacking in precision of meaning. In addition, during the parent interview, the parents' English speech tended to be more heavily accented than the children's (from the experimenter's impressions). It thus becomes more understandable that parents' restricted English speech would not improve the subject's performance on the WISC.

The last combined exposure to English variable was FT. Here the correlations with the Verbal, Performance, and Overall totals were .26, .14, and .22, respectively, and were treated as indicating no relation.

Future Prospects

There is clearly a need to replicate this study with a larger, less select sample. This study only begins the examination of cultural and class differences as they are related to mental abilities of Chinese-Americans. In addition to replicating, future studies should compare and contrast the patterns and levels of abilities using different tests than those used in the present study, with several ethnic and social class groups, as in the Lesser et al. (1965) study. Such studies necessitate a multicultural research staff. Not only must the testers but also the analyzers and interpreters of the result be sensitive to the subjects' cultural uniqueness, as in this study.

Shifting from a multicultural to a multiclass orientation, one can also begin to explore the differences in mental abilities across social class levels and the interactions of class and cultural differences as they relate to mental abilities.

After the identification of these differences, one can begin to examine possible sources of the differences. In this study of Chinese subjects, several indices of exposure to English were correlated with the Verbal, Performance, and Overall totals. These indices were composed to reflect the apparent factor-analytic unities in this ecology. Would similar composite indices emerge in other, larger studies? How could an exhaustive listing of cultural variables not found or explored in this study contribute significantly and better represent cultural and class differences? Are variables responsible for cultural differences independent of variables responsible for class differences? How should cultural definitions of class be introduced, if at all? How can the issue of genetic determination best be introduced? While the present study hardly helps to answer such questions, it does contradict the usual finding that higher social class and greater exposure to English lead to higher WISC scores in minority groups. Hence, it may help to refute an overly simplistic explanation of average group differences in WISC scores.

REFERENCES

Anastasi, A. *Differential psychology*. New York: Macmillan, 1967.

Backman, M. E. Patterns of mental abilities of adolescent males and females from different ethnic and socio-economic backgrounds. *Proceedings of the 79th Annual Convention of the American Psychological Association*, 1971, **6**, 511–512. (Summary)

Bernstein, B. Social class and linguistic development: A theory of social learning. In A. H. Halsey, J. Floud, & C. A. Anderson (Eds.), *Education, economy, and society*. New York: Free Press of Glencoe, 1961.

Burnes, K. Patterns of WISC scores for children of two socio-economic classes and races. *Child Development*, 1970, **41**, 493–499.

Christiansen, T., & Livermore, G. A comparison of Anglo-American and Spanish-American children on the WISC. *Journal of Social Psychology*, 1970, **81**, 9–14.

Curry, R. L. The effect of socio-economic status on the scholastic achievement of sixth-grade

children. *British Journal of Educational Psychology*, 1962, **32**, 46–49.

Darcy, N. T. Bilingualism and the measurement of intelligence: Review of a decade of research. *Journal of Genetic Psychology*, 1963, **103**, 259–282.

Dictionary of occupational titles. (3rd ed.) Washington, D.C.: U.S. Government Printing Office, 1965. 2 vols.

Eells, D., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. *Intelligence and cultural differences*. Chicago: University of Chicago Press, 1951.

Hess, R. D. Class and ethnic influences upon socialization. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology*. New York: Wiley, 1970.

Hess, R. D., & Shipman, V. C. Early experience and the socialization of cognitive modes in children. *Child Development*, 1965, **36**, 869–886.

Kahl, J., & Davis, J. A comparison of indexes of socio-economic status. *American Sociological Review*, 1955, **20**, 317–325.

Kohn, M. L. *Class and conformity*. Illinois: Dorsey Press, 1969.

Lesser, G. S., Fifer, G., & Clark, D. H. Mental abilities of children from different social class and cultural groups. *Monograph of the Society for Research in Child Development*, 1965, **30**, No. 102.

Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill, 1967.

Peal, E., & Lambert, W. E. The relation of bilingualism to intelligence. *Psychological Monographs*, 1962, **76** (27, Whole No. 546).

Soffietti, J. P. Bilingualism and biculturalism. *Journal of Educational Psychology*, 1955, **48**, 222–227.

(Received March 12, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 835-844

EFFECT OF HYPOTHESIS/TEST TRAINING ON READING SKILL[1]

S. JAY SAMUELS,[2] PATRICIA DAHL, AND TEARA ARCHWAMETY

University of Minnesota

A task analysis on the components underlying the hypothesis/test model of word recognition generated seven subskills. In Experiment 1, 60 retarded children were randomly assigned to an experimental and control group. Experimental subjects were taught hypothesis/test word-recognition subskills. In Experiment 2, 40 normal children were randomly assigned to a 2 × 2 factorial design. The factors were training on (*a*) hypothesis/test subskills and (*b*) recognizing high-frequency words flashed with a projector. Results of both experiments indicated hypothesis/test training produced significantly superior word recognition and comprehension. Discussion centers on the need for developing subskills and strategies to the point where they are a unitary process that can be performed automatically without attention, so as to permit rapid comprehension.

An important aspect of reading pedagogy is the discovery of skills that are important but not taught. At the present time, reading methods tend to focus on initial skills, and we need methods that train more sophisticated strategies.

Possible reasons for the dearth of research on training intermediate and fluent reading processes include the following: Reading, as a covert process, is difficult to study; descriptions of the word-recognition process appear to be in conflict (Cattell, 1885; Gough, 1972; Kolers & Lewis, 1972); and until recently we did not have explicit, testable models of recognition and information processing (Williams, 1971). The recent advances in models of recognition, reading, and information processing (Davis, 1971; Singer & Ruddell, 1970) have been most useful in suggesting explicit pedagogical approaches for teaching more sophisticated reading strategies than has heretofore been the case.

The research reported here operationalizes a method for teaching intermediate skills based on a partial model of word recognition (Samuels, 1970). The partial model of word recognition is not an entirely new idea. In many ways it is similar to the Halle and Stevens (1964) analysis by synthesis model of speech recognition and to Bruner's (1951) hypothesis testing and Solley and Murphy's (1960) trial and check. Ryan and Semmel (1969) have also written about the word-recognition process in which they state, "Expectancies about syntax and semantics within context lead the reader to form hypotheses which can be confirmed or not confirmed with only a small portion of the cues available in the text [p. 59]." What these models appear to acknowledge is that recognition is a constructive process in which output is different from and greater than the input. Implications of these models for reading would be that, given context and

[1] This research was funded by grants from the National Institutes of Child Health and Human Development to the Minnesota Reading Research Project; the Bureau of Education for the Handicapped Division of Research to the Minnesota Research, Development and Demonstration Center in Education of Handicapped Children; and the National Institute of Child Health and Human Development and National Science Foundation to the Center for Research in Human Learning.

The authors extend appreciation to the Minneapolis, Minnesota, and to the Bloomington, Minnesota schools for their help.

[2] Requests for reprints should be sent to S. Jay Samuels, Center for Research in Human Learning, University of Minnesota, Elliott Hall, Minneapolis, Minnesota 55455.

but a single letter from a target word, the reader ought to be able to generate the entire word from this minimal, visual-letter cue.

According to the partial model of word recognition, four processing stages are involved in recognizing a target word. In the first stage the words preceding the target word are read for meaning. This information is used in the second stage to generate one or more hypotheses as to the identity of the target word. In the third stage, visual information consisting of one or more letters is picked up from the target word and tested against the predicted word. In the final stage, the hypothesis is accepted or rejected depending on whether the word fragments perceived match the expected word. Speed of recognition is determined partly by the amount of visual information from the target word necessary for verifying a prediction. The less visual information required, the faster is the recognition.

Predictions based on this model suggest that skilled readers have better word recognition because of superior processing strategies at each of the four stages described above. Tests of this hypothesis comparing adults and children (Samuels & Chen, 1972) and good readers versus poor readers at the elementary school level (Begy & Samuels, in press) indicated that the better readers were more accurate in generating the target word when given context; they required less visual information from a target in order to identify the word; they were better able to identify letters from the target word which could serve as a cue to recognition (in the absence of total recognition); and they were more willing to alter an incorrect identification of a target word.

While this partial model of word recognition has proven useful, it has a major problem. The amount of time necessary to generate a prediction is in the neighborhood of 200 milliseconds (Posner & Boies, 1971). Since it takes about 250 milliseconds or less to recognize a word in isolation, this model does not account for the high-speed recognition responses of fluent readers reading meaningful material (Cosky & Gough, 1973). While the hypothesis/test procedure is too slow for what goes on in fluent reading, the model seems to account for intermediate levels of reading skill. For the high-speed recognition responses of fluent readers, the model requires further refinement.

Other researchers have tried to develop new reading methods based on the use of context for predicting a target word. Many of these experimental methods used the cloze technique but failed to find improvement in reading as a result of this practice (Bloomer et al., 1966; Blumenfield & Miller, 1966; Friedman, 1964; Guice, 1969; Schneyer, 1965). It should be noted, however, that these studies did not provide training on the subskills necessary for successful cloze performance and the failure to find differences may have been due to the students' inability to perform and profit from cloze-type exercises. Kennedy and Weener (1973), however, did find improved comprehension as a result of training on cloze passages even though they did not provide instruction in cloze subskills.

For the study described in this report, a task analysis was done on each stage of the partial model of word recognition. Instructional sequences were then designed to teach the subskills identified by the task analysis. For example, methods were designed for instructing children how to predict from context and how to use minimal visual information in verifying a prediction. Pilot work and models of speech recognition (Halle & Stevens, 1964) suggested that these subskills were already mastered in the auditory-speech mode. Consequently, instruction was sequenced to start with skills in the auditory-speech mode and to progress to the skills required for recognizing printed words.

The purpose of the two studies reported here was to test the extent to which training on the subskills underlying the partial model of word recognition would lead to improved reading. One of the dependent variables investigated was accuracy of word recognition. Another was speed of word recognition, since the model suggests that using minimal visual information should lead to increased word-recognition speed. The third

dependent variable was comprehension as measured by a modified cloze. Bormuth's (1966) research indicates that cloze tests are measures of reading comprehension. To enable subjects to use the hypothesis/test strategy of recognition, the subskills identified by the task analysis were taught to the experimental subjects. In Experiment 1, mentally retarded subjects were used. Since failure to find differences in this experiment may have been due to failure to adequately teach the subskills, tests were given to determine if the subjects had learned the subskills. In Experiment 2 a replication and extension were done using normal subjects. In addition to hypothesis/test training, the second experimental factor was training subjects to recognize high-frequency words flashed with a projector. McGinitie (1973) stated that if one knows the 1,000 high-frequency words in English, one can identify three fourths of the words used in school books. Consequently, to aid the use of context, some students in the second study were given extensive training with these basic words. Shankweiler and Liberman (1972) found that the ability to recognize words in isolation is highly related and fundamental to the ability to read in context.

Experiment 1

Method

Subjects

Sixty mentally retarded children were randomly assigned to an experimental and control group, 30 to a group. The mean IQ for the retarded subjects was 72 and the mean age was 10.3 years. Experimental subjects were trained on the subskills underlying the hypothesis/test strategy of word recognition. Control subjects were given regular reading instruction. Instructional time was carefully controlled for both groups. Training for both groups lasted 14 weeks, approximately four hours a week. Experimental subjects were taught by three undergraduate assistants. Each assistant worked with 10 retarded subjects. The undergraduates had no previous teaching experience nor did they have a background in reading pedagogy. Each week the assistants were given 10–15-minute explanations of how to teach the subskills. Control subjects were taught by three experienced certificated teachers with training in special education. Each control teacher worked with 10 subjects. To control possible Hawthorne effects, a graduate research assistant, with experience in teaching reading, gave each of the control subgroups an additional hour of reading instruction each week. The assistant had the control subjects read from their reading books and their comprehension was checked.

Design

The design was a simple randomized design.

Procedure (*Independent Variable*)

Hypothesis/test. The following subskills and methods of instruction were derived from a task analysis of the model of word recognition. (For examples of materials and methods used to teach these subskills see Archwamety & Samuels, 1973; Appendix A, Dahl, Samuels, & Archwamety, 1973.) The seven component skills derived from the model are as follows:

1. Training on the ability to say a word given an initial sound. The subject was given drills of the following nature:

Stimulus: The experimenter says, "Tell me a word starting with the sound */p/*."
Response: The subject gives a word starting with that sound. Any word starting with the sound is acceptable.

2. Training on the ability to determine the beginning letter of a spoken word. The subject was given drills of the following nature:

Stimulus: The experimenter asks, "What is the first letter in the word *girl*?"
Response: The subject gives the name of the initial letter in *girl*, that is, *g*.

3. Training on the ability to recognize visually the initial letter of a word presented orally. The subject was given drill of the following nature:

Stimulus: The experimenter says, "What is the first letter in the word *boy*?" Then the experimenter shows a card with the letters *b, c, t, r* printed on it.
Response: The subject points to the letter *b*.

4. Training on the ability to use auditory context to predict words that could logically follow. The subject was given a drill of the following nature:

Stimulus: The experimenter says, "My mother sleeps on her ——."
Response: The subject predicts the missing word. Any word that makes sense is acceptable.

5. Training on the ability to use auditory context to predict word(s) that could logically follow in a sentence hearing just the initial sound of the word. The subject was given drill of the following nature:

Stimulus: The experimenter says, "The cat ran after the /m/____."
Response: The subject predicts what the missing word might be. The words must begin with the /m/ sound and make sense in the context.

6. Training on the ability to use visual context to predict word(s) that would logically follow in a sentence without seeing the initial letter of the word. The subject was given drill of the following nature:

Stimulus: The experimenter shows the following in printed form: The children open the ____.
Response: The subject is asked to read and predict the word in the blank. The experimenter tells the subject whatever word the subject cannot read in context. Any word that makes sense is acceptable.

7. Training on the ability to use visual context to predict word(s) that could logically follow in a sentence when given the initial letter of the target word. The subject was given drill of the following nature:

Stimulus: The experimenter shows the following printed form: The girl ate the b____.
Response: The subject is asked to read this and predict the word in the blank. Any word beginning with *b* that makes sense is acceptable.

The assistants were told that accuracy in response was desirable but not sufficient. What was desired was accuracy and speed of response. These goals of accuracy and speed were emphasized during instructional sessions with the experimental subjects.

TABLE 1
WORDS TACHISTOCOPICALLY PRESENTED WITH EACH EXPERIMENTAL CONDITION

Condition	Word: S_1	Target word: S_2
1	—	pard
	—	rabbi
2	—	camp
	—	fifty
3	cold	snow
	green	grass
4	blue	ocean
	beautiful	song
5	Lemon has a salty	taste
	We heard a loud	noise

Procedure (Dependent Variables)

Speed of tachistoscopic word recognition. The two groups of subjects (experimental and control) were tested on speed of word recognition, using a two-channel, scientific prototype tachistoscope to present stimuli. Each subject was tested under five conditions. The five conditions and the words used in the conditions may be seen in Table 1. Under each of the conditions there were two target words represented by S_2 in Table 1.

Description of five conditions. The target words of Condition 1 were low-frequency words that occur less than six times in one million words (Thorndike & Lorge, 1944). The target words in other conditions were higher frequency words, but their frequency values were controlled to be approximately equal to one another, about 400–600 occurrences per million words. These frequency values are averages of the Thorndike-Lorge word-frequency count.

Words in Condition 3 and Condition 4 were selected such that the associative values between them in Condition 3 were high (47 between *cold* and *snow*, 113 between *green* and *grass*) and those between S_1s and S_2s in Condition 4 were low (0 between *blue* and *ocean*, 0 between *beautiful* and *song*). These associative values were taken from Palermo and Jenkin's (1964) word association norms.

Administration of test. Each subject was given warm-up training on the tachistoscope to familiarize him with the task. The warm-up tasks were analogous to the conditions found in the test proper, but different words were used.

The five conditions were randomly administered. The method of ascending limits was used in determining visual recognition threshold. That is, the first time the word appeared the exposure time was very short. With each subsequent presentation of the word the exposure time was increased until the subject gave two correct recognition responses. For purposes of analysis, the average visual duration threshold of the two correct responses was used.

As soon as the subject read the first S_1 word of a pair, it was terminated and then for 20 milliseconds the target word was flashed. The subject was then to identify that word. If he could not, the word pair was presented again and again, with the exposure duration of the target word increasing 10 milliseconds each time. An experimental event was completed as soon as the subject could identify the target twice.

Word identification in context—accuracy of recognition. Materials consisted of two sets of 5 × 8 inch index cards, with 10 cards in each set. On each card a single sentence was typed using an Underwood primary typewriter. The last word in the sentence was underlined in red and was considered to be the target word while all words preceding the target word were considered context. Sentences in

the first set were such that the context was so compelling that the target word could hardly be any other. An example is, "It is dark at *night*." In the second set the context was less compelling as in, for example, "There are fish in the lake."

The procedure consisted of showing the child the sentence and having the child read the sentence to the experimenter. Help was given on context words but no help was given on the target word.

Modified cloze test of comprehension. A 122-word passage with 20 deleted words was shown to the student. Each deletion contained the first letter of the target word, as exemplified in, "The boys r____ their bicycles back h____." Help was given to the students on context words, but none was given on the target. Only exact replacements were counted as correct.

Results

Seven Component Skills of Hypothesis/Test Procedure

A pretest was given on each of the subskills. Seven *t* tests were run comparing experimental and control groups on each of these component skills. Significant differences were not found between the two groups on any of the seven measures on either number of correct responses or latency of response.

At the end of the training period *t* tests were again computed for each of the component subskills. The experimental group was significantly superior ($p < .05$ versus $p < .01$) on accuracy criteria on six of the seven component skills. The only skill where no difference was found between the two groups was Subskill 4 as described in the method section. When one compares the latency of a correct response, the experimentals were faster ($p < .05$ versus $p < .01$) on five of the seven measures. Those subskills failing to reach significance were Skills 1 and 5.

With regard to the question of could these subskills be taught, the overall superiority of the experimental group on accuracy and latency criteria indicates that the method used to teach the skills was effective.

Speed of Tachistoscopic Word Recognition

A prerequisite for inclusion in the tachistoscopic study was the ability to read all the words shown with the machine. Only 20 subjects from the experimental and 20 from the control met this requirement. The data for the speed of word recognition were analyzed with a 2 × 5 analysis of variance with repeated measures on word-presentation conditions (see Table 2).

TABLE 2

Speed of Tachistoscopic Word Recognition of High-Frequency (HF) and Low-Frequency (LF) Words Presented with and without Context

Condition	Experimental group		Control group	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*
Isolation (LF)	34.25	11.54	66.00	64.74
Isolation (HF)	33.50	15.50	78.25	99.20
One-word context (H)	47.50	73.70	118.50	123.65
One-word context (L)	81.50	89.92	117.25	163.99
Sentential context	45.50	56.61	105.25	187.82
Combined	48.45	43.08	97.05	84.03

The between-subject source of variance on the analysis of variance indicated that the experimental group recognized the flashed words significantly faster than the control group ($F = 5.03$, df 1/38, $p < .05$). The within-subject source of variance on the analysis of variance indicated that there was no difference among the five word-presentation conditions ($F = 1.92$, df 4/152, *ns*). The interaction effect was not significant ($F < 1$).

Word Identification in Context

Table 3 shows the mean number of correct responses given to the target words when they were preceded by either a compelling or ambiguous context. The 2 × 2 analysis of variance with repeated measures on one of the factors indicated that overall (i.e., compelling and ambiguous context combined) the experimental group identified target words more accurately than did the controls ($F = 4.92$, df 1/58, $p < .05$). The analysis also indicated that identification was more accurate under compelling context than under ambiguous context ($F = 46.79$, df

TABLE 3
SUMMARY SCORES ON WORD IDENTIFICATION IN CONTEXT

Condition	Experimental group		Control group	
	$\bar{X}$	SD	$\bar{X}$	SD
1. Compelling context	5.47	2.57	5.43	2.03
2. Ambiguous context	4.57	2.30	2.40	1.74

1/58, $p < .01$). There was also a group by condition interaction ($F = 13.76$, df 1/58, $p < .01$). Consequently a test of simple effects was done that indicated no significant difference between experimental and controls under compelling context. However, when the target word was preceded by ambiguous context the hypothesis/test group was significantly superior at identifying the word ($F = 14.30$, df 1/116, $p < .01$).

Modified Cloze Test of Comprehension

The mean number of correct words given by the experimental group was 15.10 (SD = 2.60) and the mean number for the control group was 10.67 (SD = 4.31). The t test indicated superior performance for the experimental group ($t = 4.74$, df 58, $p < .01$).

EXPERIMENT 2

Method

Subjects

The third-grade student body in a middle-class, suburban elementary school was given the Iowa Tests of Basic Skills and 36 of the poorest readers were selected as subjects. They were randomly assigned to one of four groups. The mean IQ for the four groups was 107.00, and the mean reading achievement score was 3.05.

Design

A 2 × 2 factorial design was used. The factors were hypothesis/test training and flashed word recognition training.

Procedure (Independent Variables)

Hypothesis/test training. The same subskills and methodology were used as described in Experiment 1, method section.

Flashed word recognition. The subjects who received training recognizing flashed words were given practice in word recognition beyond accuracy. Practice was accomplished by flashing words with a carousel projector at a rate controlled by a timer. The words were first exposed at a visual duration threshold of 3 seconds and advanced to shorter exposures (2 seconds, 1.5 seconds, 1 second, .5 second). Change to faster rates was determined by going through all 80 words in a tray perfectly three times. When a subject was 100% accurate at the fastest rate, the subject selected a new tray of slides. The 1,600 different words in the carousel trays were selected from the Dale-Chall list of 3,000 most common words.

Teachers and time. The teacher who taught the experimental groups had six years of teaching experience; however, she had only 1 year of experience at the grade level. The control teacher had 19 years of teaching experience about equally divided between second and third grades. The amount of time devoted to reading was the same for experimental and control groups. The experiment was run for seven months, at the end of which time the tests were given.

Procedure (Dependent Variables)

Tachistoscopic word recognition. The exact procedure, words, and equipment were used in Experiment 2 as were described in Experiment 1.

Modified cloze test of comprehension. A 610-word passage with 62 deletions was shown to the students. Each deletion contained the first letter of the target word. As in the previous study, help was given on context words but not with target. Only exact replacements were counted.

RESULTS

To determine the initial comparability of the four randomly assigned groups, tests were run on several variables. No difference was found on The Iowa Tests of Basic Skills—Reading Comprehension; the Cognitive Abilities Intelligence Test; and a test of tachistoscopic word recognition. Although no significant differences were found, the control group that received no hypothesis/test training had faster tachistoscopic word-recognition scores on four of the five treatments.

Tachistoscopic Word Recognition

Means on speed of word recognition for the four groups can be found in Table 4.

Speed of tachistoscopic word recognition at the end of training was tested using a three-factor analysis with repeated measures on one of the factors. The two between-subject factors were hypothesis/test

versus control and carousel word recognition versus control. The within-subject factor consisted of the five conditions of tachistoscopic word recognition as shown in Table 1.

The analysis of variance on speed of word recognition used gain scores. It indicated on the between-subject source of variation that the experimental hypothesis/test trained group was significantly faster than its control ($F = 9.35$, df 1/32, $p < .01$) but there was no difference between the group trained in carousel word recognition and its control ($F < 1$, df 1/32). On the within-subject source of variation there were significant differences in speed of recognition on the five repeated measures conditions of word recognition ($F = 24.77$, df 4/128, $p < .01$). None of the interactions on either the between- or within-subject variations were significant.

TABLE 4

SPEED OF RECOGNITION IN MILLISECONDS ON TACHISTOSCOPIC TEST

Treatment	Condition				
	5[a]	4[b]	3[c]	2[d]	1[e]
Pre					
H	105	91	115	115	110
$\bar{H}$	85	85	115	107	121
Post					
H	53	61	88	90	140
$\bar{H}$	55	73	106	108	175
Change					
H	52	30	27	25	−30
$\bar{H}$	30	12	9	−1	−54
Pre					
A	105	93	128	111	124
$\bar{A}$	85	84	101	111	107
Post					
A	63	74	103	109	174
$\bar{A}$	45	59	90	90	141
Change					
A	42	19	25	2	−50
$\bar{A}$	40	25	11	21	−34

Note. H = Hypothesis/test training; $\bar{H}$ = no hypothesis/test training; A = flashed word-recognition training; $\bar{A}$ = no flashed word-recognition training.

[a] Isolated word, low frequency.
[b] Isolated word, high frequency.
[c] One word context, low association.
[d] One word context, high association.
[e] Sentence context.

TABLE 5

MEANS AND STANDARD DEVIATIONS ON CLOZE TEST ERRORS

Treatment	$\bar{X}$	SD	N
HA	11.44	3.98	9
H$\bar{A}$	9.33	2.58	9
$\bar{H}$A	16.22	4.76	9
$\bar{H}\bar{A}$	16.00	5.14	9
H	10.39	3.51	18
$\bar{H}$	16.11	4.95	18
A	13.83	4.99	18
$\bar{A}$	12.67	5.26	18

Note. H = Hypothesis/test training; $\bar{H}$ = no hypothesis/test training; A = flashed word-recognition training; $\bar{A}$ = no flashed word-recognition training.

Modified Cloze Test of Comprehension

Means and standard deviations on the cloze test errors can be found in Table 5.

The two-way analysis of variance indicated that the hypothesis/test trained groups made significantly fewer errors than the controls ($F = 14.65$, df 1/32, $p < .01$). On the second factor, no difference was found between the groups given carousel word-recognition training and the controls ($F < 1$, df 1/32). The interaction effect was not significant.

DISCUSSION

A partial model of word recognition was transformed by means of a task analysis into a series of subskills. This model uses hypothesis/test procedures for word recognition. The purpose of the two studies reported in this investigation was to determine if children who were trained in these hypothesis/test subskills would be superior

in word recognition and comprehension. An additional question investigated in Experiment 1 was whether the subskills could be taught effectively.

Results of Experiment 1 indicated that the subskills could be taught effectively, in that the mentally retarded children who were in the experimental group were generally superior to their controls in both accuracy and response latency on the tests of subskill mastery. Of greater concern was the performance of the experimentally trained subjects on the two tests of word recognition. One test consisted of recognizing words presented with a tachistoscope. On this test the experimentally trained group was significantly faster in speed of word recognition. The second test of word recognition had two parts. In the first part the target word was preceded by a compelling context, whereas in the second part the target word was preceded by ambiguous context. The analysis indicated no difference between experimental and control groups on the compelling context but significantly superior performance by the experimental group on the ambiguous context, where preceding context did not readily elicit the target word. The cloze test of comprehension indicated the superiority of the hypothesis/test trained experimental subjects.

Results of Experiment 2, which used a normal population, were much the same. On the test of tachistoscopic word recognition, the experimentally trained subjects were significantly faster, when one compared change in performance from pretests to posttests. The cloze test of comprehension indicated the superiority of the experimental subjects.

It should be pointed out that in Experiment 2 flashed word recognition was added to the design as a second factor. The rationale for including this factor is that in order to be able to use context as an aid in recognizing a target word, the reader must be able to recognize the words in context accurately and quickly. Training on this variable consisted of recognizing high-frequency words that were flashed with a projector. Comparing experimental and control groups on this factor indicated no difference between them on tests of word recognition or comprehension.

Previous portions of this article outlined the rationale for teaching the hypothesis/test word-recognition skill and the component subskills. At this point it is necessary to mention that the goal in teaching this word-recognition strategy is for each of the subskills to be integrated into a higher-order, wholistic process. Hilgard and Marquis (1961) wrote:

> most learning is complex and requires the simultaneous learning of several components. Possibly the clearest illustration of this is the appearance during learning of plateaus (Bryan & Harter, 1897) during which practice does not lead to improvement. In complex learning, plateaus often seem to be temporary periods devoted to the organization of component habits into larger units [p. 127]....

There are numerous psychomotor skills that, through extensive training, appear to be a single skill. The tennis serve, for example, appears to be one fluid motion, when, in fact, in teaching the skill the coach breaks it into its components. Recent research by Guthrie (1973) in the area of reading indicates that for good readers, reading is a unitary process. However, for poor readers the process is not unitary and consists of a number of independent subskills. Guthrie's research implies that in the process of becoming a skilled reader, the subskills must be integrated in such a way that they become unified into a single process that we call "reading."

As necessary as the integration of these subskills appears to be, there is still an additional requirement for the development of a good reader, namely, that the skills used in reading be run off with no attention. When attention is required for the execution of a skill, immediate access of meaning is prevented (LaBerge & Samuels, 1974). Fluent readers are characterized by their ability to perform the necessary decoding automatically—without attention—so that their limited attention and memory capacity is left free to process meaning. At the present time, the psychology of reading has dealt

with the problems inherent in developing accuracy in word recognition as well as the delineation of some of the subskills involved in the reading process. The field has not yet dealt sufficiently with the problem of efficient training strategies for going beyond accuracy to automaticity, so that Huey's (1908) advice about how to develop automaticity is still the level at which we are operating today: "repetition progressively frees the mind from attention to details, makes facile the total act, shortens the time, and reduces the extent to which consciousness must concern itself with the process [p. 104]."

In order for the subskills to be executed as a wholistic process without the need for attention, an extended amount of practice beyond mere accuracy is required.

REFERENCES

Archwamety, T., & Samuels, S. J. *A mastery based experimental program for teaching mentally retarded children word recognition and reading comprehension skills through use of hypothesis/test procedures.* (Tech. Rep. No. 12, University of Minnesota Reading Research Project, Grant No. OE-09-332189-4533 [032]). Washington, D.C.: Bureau of Education for the Handicapped, United States Office of Education, 1973.

Begy, G. L., & Samuels, S. J. *Comparison of word recognition strategies for good and poor readers.* (Occasional Paper) University of Minnesota, Research, Development & Demonstration Center in Education of Handicapped Children, 1974.

Bloomer, R., et al. *Non-overt reinforced cloze procedures* (Cooperative Research Project 2245, University of Connecticut). Washington, D.C.: United States Office of Education, 1966.

Blumenfield, J., & Miller, G. Improving reading through teaching grammatical constraints. *Elementary English,* 1966, **43,** 752–755.

Bormuth, J. R. Readability: A new approach. *Reading Research Quarterly,* 1966, **1,** 79–132.

Bruner, J. S. Personality dynamics and the process of perceiving. In R. R. Blake & G. V. Ramsey (Eds.), *Perception: An approach to personality.* New York: Ronald, 1951.

Bryan, W. L., & Harter, N. Studies in the physiology and psychology of the telegraphic language. *Psychological Review,* 1897, **4,** 27–53.

Cattell, J. Ueber die Zeit der Erkennung und Benennung von Schriftzeichen, Bildern und Farben. *Philosophische Studien,* 1885, **2,** 635–650.

Cosky, M., & Cough, P. B. The effect of context on word recognition. Paper presented at the Midwest Psychological Association meeting, Chicago, May 1973.

Dahl, P. J., Samuels, S. J., & Archwamety, T. *A mastery based experimental program for teaching poor readers high speed word recognition skills.* (Research Rep. 55, University of Minnesota Reading Research Project, Grant No. OE-09-332189-4533 [032]). Washington, D.C.: Bureau of Education for the Handicapped, United States Office of Education, 1973.

Davis, F. B. *The literature of research in reading with emphasis on models.* New Brunswick, N. J.: Iris Corporation, 1971.

Friedman, M. The use of the cloze procedure for improving the reading of foreign students at the University of Florida. Unpublished doctoral dissertation, University of Florida, 1964.

Gough, P. B. One second of reading. In J. F. Kavanaugh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading.* Cambridge, Mass.: MIT Press, 1972.

Guice, B. The use of the cloze procedure for improving comprehension of college students. *Journal of Reading Behavior,* 1969, **1,** 81–92.

Guthrie, J. T. Models of reading and reading disability. *Journal of Educational Psychology,* 1973, **65,** 9–18.

Halle, M., & Stevens, K. N. Speech recognition: A model and a program for research. In J. A. Foder & J. J. Katz (Eds.), *The structure of language: Readings in the philosophy of language.* Englewood Cliffs, N. J.: Prentice-Hall, 1964.

Hilgard, E. R., & Marquis, D. G. *Conditioning and learning.* New York: Appleton-Century-Crofts, 1961.

Huey, E. B. *The psychology and pedagogy of reading.* Cambridge, Mass.: MIT Press, 1908.

Kennedy, D. K., & Weener, P. Visual and auditory training with the cloze procedure to improve reading and listening comprehension. *Reading Research Quarterly,* 1973, **8,** 524–541.

Kolers, P. A., & Lewis, C. L. Bounding of letter sequences and the integration of visually presented words. *Acta Psychologica,* 1972, **36,** 112–124.

LaBerge, D., & Samuels, S. J. Toward a theory of automatic information processing in reading. *Cognitive Psychology,* 1974, **6,** 293–323.

McGinitie, W. H. Testing reading achievement in urban schools. *The Reading Teacher,* 1973, **27,** 13–21.

Palermo, D. S., & Jenkins, J. J. *Word association norms: Grade school through college.* Minneapolis, Minn.: University of Minnesota Press, 1964.

Posner, N., & Boies, S. Components of attention. *Psychological Review,* 1971, **78,** 391–408.

Ryan, E. B., & Semmel, M. I. Reading as a constructive language process. *Reading Research Quarterly,* 1969, **5,** 59–83.

Samuels, S. J. Recognition of flashed words by children. *Child Development,* 1970, **41,** 1089–1094.

Samuels, S. J., & Chen, C. Comparison of word recognition strategies of adults and children. In *Reading: Process and pedagogy, Twentieth yearbook of the national reading conference*. Milwaukee, Wis.: Marquette University, National Reading Conference, Inc., Reading Center, 1972.

Schneyer, J. Use of the cloze procedure for improving reading comprehension. *The Reading Teacher*, 1965, **19**, 174–179.

Shankweiler, D., & Liberman, I. Y. Misreading: A search for causes. In J. F. Kavanaugh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading*. Cambridge, Mass.: MIT Press, 1972.

Singer, H., & Ruddell, R. B. (Eds.) *Theoretical models and processes of reading*. Newark, Del.: International Reading Association, 1970.

Solley, C. M., & Murphy, G. M. *The development of the perceptual world*. New York: Basic Books, 1960.

Thorndike, E. L., & Lorge, I. *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University, 1944.

Williams, J. P. Learning to read: A review of theories and models. In F. B. Davis (Ed.), *The literature of research in reading with emphasis on models*. New Brunswick, N. J.: Iris Corporation, 1971.

(Received February 25, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 845–850

SELF-DETERMINATION OF ACADEMIC STANDARDS BY CHILDREN:

TOWARD FREEDOM FROM EXTERNAL CONTROL[1]

JEFFREY J. FELIXBROD AND K. DANIEL O'LEARY[2]

State University of New York at Stony Brook

The effects of contingent reinforcement, with performance standards self-determined or externally imposed, upon children's productivity in a naturalistic setting were examined. Children in one contingent reinforcement condition self-determined their academic performance standards, while the same requirements were externally imposed upon yoked children in a second contingent reinforcement condition. Those in a no-reinforcement control condition performed in the absence of external reward. In a reinforcement phase, children in the self-determination condition were significantly more productive than control subjects and performed as well as those for whom academic standards were externally imposed; however, when reinforcement contingencies were later absent for all groups, there was a tendency, not statistically significant, for children in the self-determination condition to be less productive than those who had experienced externally imposed standards.

There have been countless demonstrations that children's behavioral productivity can be increased through contingencies administered and controlled by an outside agent. This line of investigation recently has been cross-fertilized by studies of behavioral self-control. A number of experiments have disclosed that behavioral self-management may serve as a potential effective alternative to external control (e.g., Bandura & Perloff, 1967; Bolstad & Johnson, 1972; Drabman, Spitalnik, & O'Leary, 1973; Kaufman & O'Leary, 1972). However, although these studies have examined components of behavioral self-regulation such as self-instruction, self-monitoring, and self-administration of reinforcement, little research has focused upon a comparison of the effects of contingent reinforcement in situations in which performance standards (i.e., behavioral requirements for differential levels of reward) are self-determined rather than externally imposed. The few investigations which have examined the effects of contingent reinforcement under conditions of self-determined and externally imposed performance standards have generated contradictory findings. Lovitt and Curtiss (1969) found that the academic response rate of a 12-year-old boy was higher when he, rather than the teacher, selected his educational requirements. On the other hand, Glynn (1970) observed no difference

[1] This study was supported by Grant OEG-0-71-2872 from the U.S. Office of Education. The project reported herein was performed pursuant to a grant from the U.S. Office of Education, Department of Health, Education, and Welfare. The opinions expressed herein, however, do not necessarily reflect the position or policy of the U.S. Office of Education, and no official endorsement by the U.S. Office of Education should be inferred. Thanks are due Richard Hoff, principal, and the staff of the Wing Street Elementary School, Middle Country School District, Long Island, New York. The dedication of the experimenters, Diane Jillie, Aileen Leventon, Michele Domash, and Linda Duberstein, was especially appreciated. Robert M. Liebert provided sound advice in every phase of the project, and Ronald Kent offered helpful suggestions concerning data analysis.

[2] Reprints may be obtained from Jeffrey J. Felixbrod or K. Daniel O'Leary, Department of Psychology, State University of New York, Stony Brook, New York 11790.

in academic performance as a function of student-determined and experimenter-determined reinforcement.

In a study by Felixbrod and O'Leary (1973), elementary school children in one contingent reinforcement condition self-determined their academic performance requirements for earning points which could later be exchanged for prizes. In a second contingent reinforcement condition, the same standards were externally imposed; these children were individually yoked to subjects in the first condition so that the same performance standards were employed. The two reinforcement treatments were found to be equally effective; both groups were observed to be more productive than was a no-reinforcement control group.

The significance of such findings is not fully understood, though, until the effects of withdrawing such contingencies, as well as introducing them, are known. Thus, the present experiment was designed primarily to compare the relative effects of contingent reinforcement, with amount self-determined or externally imposed, on the resistance to extinction of behavioral productivity following the withdrawal of all material incentives. Embodied in the design was a constructive replication of Felixbrod and O'Leary (1973).

It is likely, of course, that some loss of productivity will follow the withdrawal of self-determined or externally determined contingencies; critical questions include both the magnitude and rate of this decrement, relative both to a continually performing no-incentive control group and to one another. Compared to the performance of the self-determination condition in extinction, children in the external imposition condition conceivably might be more, less, or equally productive.

Method

Subjects

The subjects were 24 white children, 12 boys and 12 girls, between eight and nine years of age, drawn from two third-grade classrooms of a public school serving a lower-middle-class area of Long Island, New York. The experiment was conducted during the spring semester of the school year.

Setting and Material

The experiment was conducted in separate rooms provided by the school, where subjects were tested individually. At the beginning of every session, the child was given a packet of 12 sheets (8½ × 11 inches), each containing 20 grade-level-appropriate arithmetic problems. All subjects were presented with the same arithmetic questions in the same order during any one session. Difficulty level of the problems was constant across sessions.

Procedure

A 3 × 2 × 5 (Treatments × Experimental Phases × Sessions) factorial design was employed. Experimental phases consisted of reinforcement and extinction; there were five sessions within each experimental phase. Sessions, each of which lasted up to 20 minutes, were conducted every Monday, Wednesday, and Friday, with the exceptions of school holidays and pupil absences. In order to prevent potential confounding due to differential absence from classroom activities, subjects in all three experimental conditions were tested during the same instructional periods.

Three experimental treatment conditions consisting of 8 subjects each were formed: (*a*) a contingent reinforcement condition in which performance standards were *self-determined;* (*b*) a contingent reinforcement condition in which performance standards were *externally imposed;* and (*c*) a *no-reinforcement* control condition. The 24 children were divided into eight sets of 3 subjects, each set matched on the basis of sex and score level on two 20-minute arithmetic pretests. The 3 children within each set were randomly assigned to differing experimental conditions. Two sets of subjects then were randomly but permanently assigned to each of four white female undergraduate experimenters.

Reinforcement phase. The following instructions were read to subjects in the self-determination condition at the beginning of each of the five sessions:

1. When people work on a job, they get paid for what they do. I am going to pay you *points* which you can use to buy these prizes (pointing to prizes and point-exchange values). *Your* job is to answer these arithmetic questions. Answer the questions in order. In order to earn the points, only correct answers will count (repeated). You will have 20 minutes to do these. But you can stop before 20 minutes are up if you want to. Just come outside the door where I'll be waiting, if you want to stop sooner.

2. I am going to let *you* decide how many points you want to get paid for each right answer. Take a look at the numbers on the next page (pointing to a separate page on which the subject is to chose a performance standard). I want *you* to decide how many points you want to get paid for each right answer. (Experimenter

points to each possible choice in a list of 10 possible performance standards: "I want to get paid 1 point; 2 points; ... 10 points for each right answer.") After I leave the room, draw a circle around the number of points you want to get paid for each right answer.

3. Remember, you can stop and leave when you want to, or else I'll tell you when the 20 minutes are up. I will be right outside. (Instructions repeated if necessary.)

At the start of Sessions 2 through 5 of the reinforcement phase, the experimenter again read the instructions. During these sessions, subjects in the self-determination condition were given the opportunity to choose the same or a different performance standard.

Instructions given to children in the external imposition condition were identical to those presented to the self-determination children, except that in place of the second paragraph was the statement, "You will be paid ____ points for each right answer." In other words, the performance standards adopted by a child in the self-determination condition were applied to the subject paired with him in the external imposition condition. For example, if a particular child in the self-determination condition chose to receive 7 points per correct answer in Session 1, 9 points per correct answer in Session 2, and 10 points per correct answer in Sessions 3, 4, and 5 of the reinforcement phase, the same pattern of performance standards was set for his matched counterpart in the external imposition condition. For children in the no-reinforcement condition, instructions differed from those given to self-determination subjects in that Paragraph 2 was omitted, as were the first, second, and fifth sentences of Paragraph 1; in addition, no-reinforcement subjects were told, "You will receive no points or prizes for your work." Children in the nonreinforcement condition received the same instructions in both the reinforcement and extinction phases.

A number of procedures designed to minimize extraneous social influence were employed. Subjects performed the task in the absence of the experimenter. Children in the self-determination condition did not select a performance standard until after the experimenter had left the room. Since incidental modeling cues may influence performance in novel situations (as noted by Liebert, Spiegler, & Hall, 1970), the experimenter pointed to each of the 10 possible performance standards (rather than to only 1 or 2) that a child in the self-determination condition could select. Rewards consisted of candy and toys that ranged in value from two cents to about one dollar. In the reinforcement phase, prizes were openly displayed while children in the self-determination and external imposition conditions performed the task. No prizes were present when subjects in the no-reinforcement condition performed. The experimenters reported that, in general, the children seemed to be very impressed by the more expensive prizes. At the end of each reinforcement session, subjects in the self-determination and external imposition conditions were required to exchange for rewards all the points they had earned. Prizes were placed in paper bags which were delivered by the experimenter to the teacher who dispensed these rewards to the children at the end of the school day.

Extinction phase. In the five extinction sessions, instructions to subjects in the self-determination and external imposition conditions were identical to those given to the no-reinforcement subjects. The children were again asked to solve arithmetic problems but were informed that they would be paid neither points nor prizes for their work.

Results

Major dependent variables in this experiment were the number of correct problem solutions and the amount of time spent in the task setting; in addition, performance standards selected by self-determination subjects at the start of each reinforcement session were recorded. Reinforcement and extinction data were analyzed separately.

Correct Solutions

Figure 1 presents the mean number of problems solved correctly by the children.

Reinforcement phase. A 3 × 5 analysis of variance disclosed both a significant main effect for sessions ($F = 3.27$, $df = 4/84$, $p < .02$) and a significant Treatments × Ses-

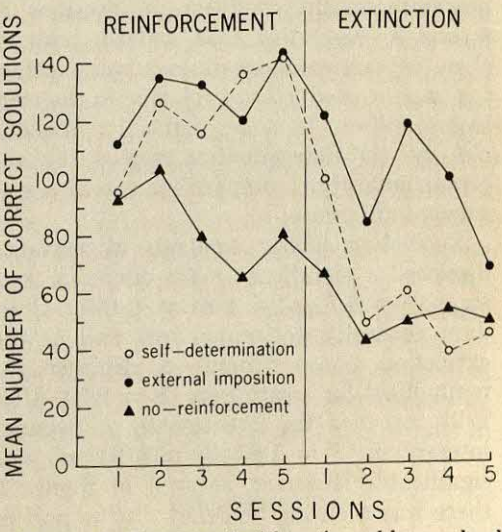


FIGURE 1. Number of arithmetic problems solved correctly by subjects in the self-determination, external imposition, and no-reinforcement conditions.

TABLE 1
NEWMAN-KEULS COMPARISONS OF MEAN NUMBER OF CORRECT PROBLEM SOLUTIONS

Reinforcement session				
1	2	3	4	5
Self-determination vs. no reinforcement				
ns	*ns*	*ns*	$p < .01$	$p < .01$
External imposition vs. no reinforcement				
ns	*ns*	$p < .01$	$p < .01$	$p < .01$
Self-determination vs. external imposition				
ns	*ns*	*ns*	*ns*	*ns*

sions interaction ($F = 2.58$, $df = 8/84$, $p < .02$); the main effect for treatments did not achieve statistical significance ($F = 2.49$, $df = 2/21$, $p < .15$). Newman-Keuls comparisons were employed to further examine the significant interaction. Table 1 presents the results of these comparisons for each of the five sessions of the reinforcement phase. The superiority in performance of the self-determination condition relative to the no-reinforcement condition was apparent in Sessions 4 and 5, while the external imposition condition was superior to the no-reinforcement condition in Sessions 3, 4, and 5, suggesting that, indeed, both of these experimental treatments were fostering greater productivity relative to the control condition. In none of the five sessions did the self-determination versus the external imposition comparisons reveal a significant difference.

Extinction phase. Analysis of variance disclosed a significant main effect for sessions ($F = 8.78$, $df = 4/84$, $p < .001$). Children generally performed less well as the extinction phase progressed. Neither the main effect for treatments ($F = 2.20$, $df = 2/21$, *ns*) nor the Treatments × Sessions interaction ($F = 1.87$, $df = 8/84$, *ns*) was significant. However, as seen in Figure 1, there was a strong tendency on the part of external imposition subjects to outperform those in both the self-determination and no-reinforcement conditions.

Time at Task

Figure 2 presents the mean number of minutes spent in the task setting.

Reinforcement phase. Both the main effect for treatments ($F = 4.92$, $df = 2/21$, $p < .02$) and the main effect for sessions ($F = 2.67$, $df = 4/84$, $p < .05$) were significant; the Treatments × Sessions interaction ($F = 1.36$, $df = 8/84$, *ns*) was not statistically significant. Individual comparisons disclosed that both self-determination and external imposition subjects spent more time in the task setting than no-reinforcement subjects ($t = 2.12$, $p < .05$, one-tailed, and $t = 2.82$, $p < .01$, one-tailed, respectively), while self-determination and external imposition subjects did not differ in time at task ($t = .50$, *ns*, two-tailed). It is apparent that the effect of reinforcement was to maintain time spent in the task setting; in the absence of reinforcement, time at task decreased.

Extinction phase. The main effect for sessions was significant ($F = 9.15$, $df = 4/84$, $p < .001$). Children generally spent less

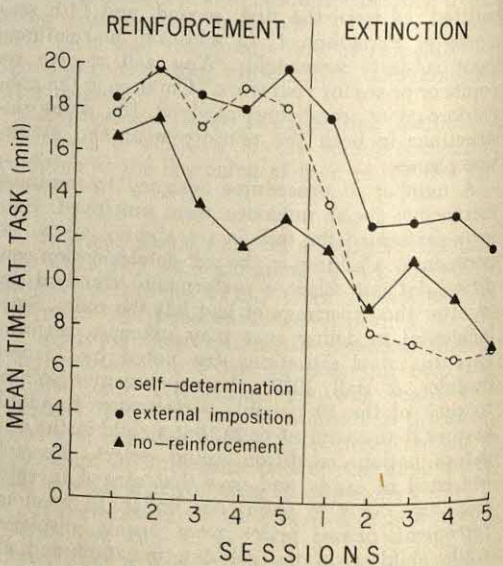


FIGURE 2. Time spent in task setting by subjects in the self-determination, external imposition, and no-reinforcement conditions.

time in the task setting as the extinction phase progressed. As was the case on the dependent measure of correct solutions, neither the main effect for treatments ($F = 1.51$, $df = 2/21$, *ns*) nor the Treatments × Sessions interaction ($F = .90$, $df = 8/84$, *ns*) was significant. However, as seen in Figure 2, there was a tendency on the part of external imposition subjects to spend more time in the task setting than those in the self-determination and no-reinforcement conditions.

Further statistical analysis revealed no sex differences in children's responses to experimental treatments.

Self-Determined Performance Standards

In the reinforcement phase, children in the self-determination condition could choose to receive between 1 and 10 points per correct solution. At the start of the first session, seven of the eight subjects selected the most lenient performance standard (10 points per correct solution), while one child chose to receive 7 points per correct solution. At the start of Sessions 2 through 5, every child selected the most lenient performance requirement. It is clear that children responded in a manner that would maximize their contingent reward.

Discussion

The overall results of the present investigation provide evidence for the efficacy of contingent reinforcement when academic standards are self-determined or externally imposed. In the reinforcement phase, contingent consequences maintained children's academic behavior equally well under conditions of self-imposed and externally imposed standards; thus, the findings of Felixbrod and O'Leary (1973) were replicated. In the extinction phase, analysis of variance disclosed no statistically significant differences between the groups on the measures of correct solutions and time at task; however, a strong tendency on the part of external imposition subjects to outperform those in the self-determination condition was noted. This latter finding is deserving of further investigation.

Children in the self-determination condition were observed to maximize their rewards by selecting the most lenient performance standard in the reinforcement phase, a finding consistent with those of Felixbrod and O'Leary (1973) and Santogrossi, O'Leary, Romanczyk, and Kaufman (1973). Apparently, many children, in the absence of social surveillance and in the presence of high magnitude rewards, will attempt to maximize their positive outcomes, especially if they can discriminate that no externally administered aversive consequences follow the self-imposition of lenient standards.

In this time of great concern over the possible misuse of behavioral technology, the development of self-management skills is a step toward freedom from external control. The results of this experiment and others (e.g., Bolstad & Johnson, 1972; Drabman et al., 1973; Felixbrod & O'Leary, 1973; Glynn, 1970) demonstrate the utility of self-management as an alternative to external control. However, in light of the present finding of self-imposed leniency in the absence of social surveillance, the task remains to determine how children could be taught to maintain reasonably stringent standards over time (i.e., do a fair amount of work for very little reward) without the need for continuous external monitoring. One method worthy of study involves the use of intermittent verbal feedback in which praise for adherence to moderately stringent self-imposed demands might be employed alone or used in combination with disapproval of self-imposed leniency. In addition, research is needed to determine the variables in a child's social learning history which may differentially affect responding during and following the administration of reinforcement under conditions of self-determined and externally imposed contingencies. Functional relationships between controlling stimuli and the behavioral components of self-control require further examination in a variety of situations.

References

Bandura, A., & Perloff, B. Relative efficacy of self-monitored and externally imposed reinforcement

systems. *Journal of Personality and Social Psychology*, 1967, **7**, 111–116.

Bolstad, O. D., & Johnson, S. M. Self-regulation in the modification of disruptive classroom behavior. *Journal of Applied Behavior Analysis*, 1972, **5**, 443–454.

Drabman, R. S., Spitalnik, R., & O'Leary, K. D. Teaching self-control to disruptive children. *Journal of Abnormal Psychology*, 1973, **82**, 10–16.

Felixbrod, J. J., & O'Leary, K. D. Effects of reinforcement on children's academic behavior as a function of self-determined and externally imposed contingencies. *Journal of Applied Behavior Analysis*, 1973, **6**, 241–250.

Glynn, E. L. Classroom applications of self-determined reinforcement. *Journal of Applied Behavior Analysis*, 1970, **3**, 123–132.

Kaufman, K. F., & O'Leary, K. D. Reward, cost, and self-evaluation procedures for disruptive adolescents in a psychiatric hospital school. *Journal of Applied Behavior Analysis*, 1972, **5**, 293–309.

Liebert, R. M., Spiegler, M. D., & Hall, M. Effects of the value of contingent self-administered and noncontingent externally imposed reward on children's behavioral productivity. *Psychonomic Science*, 1970, **18**, 245–246.

Lovitt, T. C., & Curtiss, K. A. Academic response rate as a function of teacher- and self-imposed contingencies. *Journal of Applied Behavior Analysis*, 1969, **2**, 49–53.

Santogrossi, D. A., O'Leary, K. D., Romanczyk, R. G., & Kaufman, K. F. Self-evaluation by adolescents in a psychiatric hospital school token program. *Journal of Applied Behavior Analysis*, 1973, **6**, 277–287.

(Received November 2, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 851-856

EFFECT OF PRIOR TEACHING EVALUATIONS AND LECTURE PRESENTATION ON RATINGS OF TEACHING PERFORMANCE[1]

RAYMOND P. PERRY,[2] R. RONALD NIEMI, AND KEITH JONES

University of Manitoba, Winnipeg, Canada

The study investigated the effect of prior student teaching evaluations and lecture presentation on ratings of teaching performance. One half of the subjects were introduced to a professor using a written biographical description including a positive teaching evaluation, while with the other subjects a negative evaluation replaced the positive one. Following the lecture, subjects completed a teaching evaluation. The two lectures differed only in the material presented. Results showed that prior expectations of teaching performance influenced ratings of the professor on a number of dimensions, with the positive evaluation group rating the professor significantly higher than the negative evaluation group. The effect persisted in both lectures and does not appear to be restricted to specific course material.

Various groups within universities have adopted the practice of evaluating professors' teaching performance, in part to provide feedback to the professors for possible self-improvement. Learning psychologists attest to the importance of feedback in the acquisition and change of various behavior patterns. Skinner (1957) has suggested that performance feedback acts as a discriminative stimulus to an organism, informing it what to do and what not to do in order to attain a specified goal. If effective teaching is a goal, then feedback about performance can facilitate acquisition of the necessary behaviors.

It is less certain, however, whether publicizing such information to the entire academic community is essential for improved teaching performance. Ostensibly, professor/course evaluations are made available to the general university population to inform students which professors are the best teachers. The present study considered the implications of such procedures by investigating the impact that preexisting, positive and negative evaluations have on ratings of teaching performance. It was hypothesized that the evaluations create certain expectations in judges which bias subsequent ratings in the direction of the earlier evaluations.

Empirical support for the hypothesis comes from several areas of psychology. For instance, Asch (1946) found that certain kinds of information can significantly alter ensuing personality judgments. Similar results were reported by Kelley (1950), who introduced a lecturer using a written biographical description in which the word *warm* was used to describe the lecturer's personality for one half of the subjects, while for the remaining half, the word *cold* was used. Alternating certain kinds of information caused drastic changes in personality evaluations, even though the lecture presentation was exactly the same for both groups. Rosenthal and Jacobson (1968) have shown that an individual's expectations regarding another person's behavior may provide the basis for a self-fulfilling prophecy. Expectancy effects have been found even to influence experimental outcomes in scientific research where an experimenter can obtain support for his hypothesis by producing demand characteristics

[1] Parts of this experiment were presented at the Canadian Psychological Association's Annual Convention in Victoria, British Columbia, June 1973. The authors would like to thank W. J. McKeachie for his critical comments of an earlier draft.

[2] Requests for reprints should be sent to Raymond P. Perry, Department of Psychology, University of Manitoba, Winnipeg, Manitoba R3T 2N2 Canada.

which cause the subject to respond in the predicted direction (Adair, 1973; Orne, 1962; Rosenthal, 1966).

Taken together, the research suggests that prior expectations significantly influence later evaluation. In essence, a kind of reverse Rosenthal effect was predicted in that the students' expectations were predicted to influence their future evaluations of the professor. Those students receiving the negative teaching evaluation would rate the professor's teaching performance more negatively than students who received a positive evaluation. The manipulations were carried out under two separate lecture conditions in order to assess the generalizability of the effect.

Method

Subjects

The subjects were 238 students enrolled in the introductory psychology courses at the University of Manitoba, Winnipeg, Canada. The 50 subjects who rated the 16 professor evaluations and the 103 Lecture 2 subjects participated for experimental credits as part of the introductory course. Lecture 1 subjects participated under actual classroom conditions.

Materials

The experiment consisted of two parts which differed only in the lecture material: Lecture 1 involved a brief description of several concepts in Freud's theory of personality, while the Lecture 2 material focused on some basic concepts in developmental psychology. The University of Manitoba Student Union (UMSU) anticalendar was used to select the positive and negative teaching evaluations. The UMSU anticalendar listed each professor followed by the mean ratings of each question on the UMSU professor/course evaluation questionnaire and a brief written summary of the ratings and of students' comments. The positive and negative evaluation summaries selected were as follows:

Professor D(+). As a researcher he has investigated several variables related to perceptual distortion. Primarily this research has examined various components of the visual system in humans. As a teacher, the UMSU teaching evaluations rated Professor D

> at or near the top of the academic department in all important teaching categories. Students praised the lectures given by Professor D. They were informative, lively, and stimulating. A relaxed atmosphere in the class encouraged discussion. A friendly attitude between student and professor enhanced the learning experience. His sections received one of the top course ratings in the department. The subject matter was very interesting to the students, and as one student commented, "this course is one of the most relevant I have taken." The course material was personally relevant to the students. Professor D was most highly recommended.

Professor A(−). As a researcher he has investigated several variables related to perceptual distortion. Primarily this research has examined various components of the visual system in humans. As a teacher, the UMSU teaching evaluations rated Professor A

> below average in most categories when compared to other instructors of the department. Professor A did not show much enthusiasm for teaching, did not greatly encourage class participation or outside consultation, and did not take much of a personal interest in most of the students. Lectures were boring and monotonous, and often consisted of a dictation of notes. Also, Professor A did not cover all the material in class, leaving it up to the students. In class, Professor A attempted to maintain "high school discipline."

The UMSU teacher rating form consisted of eight questions: five pertaining to the skill of the instructor; two pertaining to the interaction of the instructor with students, and one question concerning an overall rating of the professor in comparison to other university professors. Each question was followed by a five-category scale ranging from extremely unfavorable to extremely favorable. These questions are similar to those found on the McKeachie-Lin Teacher Rating Form which pinpoints six stable factors related to teaching evaluation (Isaacson, McKeachie, & Milholland, 1963).

Procedure

The teaching evaluations were taken from the 1971–1972 UMSU anticalendar which is made available to students at the beginning of the academic year. Eight positive and eight negative evaluations were selected from one discipline, psychology, and modified such that an alphabetical letter replaced the professor's actual name. The 16 evaluations were combined into a booklet and each was followed by a 9-point rating scale, ranging from extremely negative (−4) to extremely positive (+4). These booklets were administered to a group of 50 judges who rated each evaluation on the scale provided. The most extreme positive and extreme negative evaluations were selected for inclusion in the personal description of the lecturer. The summary included biographical information such as occupation and education, and this was followed by either the positive or the negative teaching evaluation (see above, Professors D and A).

A professor, blind to the experimental condi-

TABLE 1

UNWEIGHTED MEANS ANALYSES OF VARIANCE FOR EVALUATIONS OF PROFESSOR SKILL, INTERACTION, AND OVERALL EXCELLENCE

Evaluation[a]	*SS*	*df*	*MS*	*F*
1. Skill of professor				
Lecture (A)	18.60	1	18.60	4.62*
Positive/negative (B)+	1.86	1	1.86	<1
A × B	.93	1	.93	<1
Within cell	740.78	184	4.03	
2. Interaction of professor				
Lecture (A)	13.02	1	13.02	17.86**
Positive/negative (B)+	5.12	1	5.12	7.02**
A × B	.00	1	.00	<1
Within cell	134.12	184	.73	
3. Overall excellence of professor				
Lecture (A)	32.56	1	32.56	58.03**
Positive/negative (B)+	14.42	1	14.42	25.70**
A × B	1.86	1	1.86	3.32
Within cell	103.27	184	.56	

Note. These are three separate analyses.

[a] Positive/negative refers to the positive/negative teaching evaluation contained in the professor description.

*$p < .05$.

**$p < .01$.

tions, presented a lecture to a group of subjects which was divided into two parts: one half received the summary with the positive evaluation while the second half received the negative evaluation summary. In Lecture 1, the material was presented during a regular class of introductory psychology. The subjects were told only that they would be having a guest lecturer whose background was outlined in the written summary. Lecture 1 focused on fewer concepts, was concerned with an inherently more interesting topic (psychoanalysis), was intended to be more humorous and relevant, and took place during a regular class meeting. An express attempt was made to disguise the experiment, to allay any demand characteristics and, more importantly, to determine the generalizability of the phenomenon to actual classroom conditions. Subsequent questioning during debriefing indicated that the students were unaware of the manipulations while the experiment was in progress.

In Lecture 2, the subjects were volunteers from the psychology department subject pool who were told that the researchers were interested in learning and teaching effectiveness. Lecture 2 was an overview of numerous developmental concepts, dealt with more theoretical issues, made little effort to engender humor or relevancy in the issues, and took place in a laboratory setting. The biographical information and the professor were identical in both lecture conditions. Following the lecture, the subjects were administered the teaching questionnaire.

RESULTS

Three 2 × 2 factorial analyses of variance for unweighted means tested the effects of positive/negative teaching evaluation and lecture presentation on each of the following dependent measures: skill of professor, interaction of professor, and overall excellence of professor in comparison to other university professors (Table 1). A student's rating of skill of the professor was a mean of the five related questions; interaction of the professor was a mean of the two related questions; and professor excellence was based on a single question (see Table 2 for cell means). While the positive/negative evaluations did not affect the students' ratings of the professor's teaching skills ($F < 1$, $df = 1/184$, $p > .05$), they did significantly influence ratings of interaction capacity ($F = 7.02$, $df = 1/184$, $p < .01$) and ratings of overall excellence ($F = 25.70$, $df = 1/184$, $p < .01$). Students receiving the positive evaluation, as compared to students receiving the negative evaluation, rated the professor as being more oriented to student interaction (2.47 versus 2.14, respectively) and as being a better professor overall (3.97 versus 3.41, respectively) on a scale ranging from 1 to 5. It is worth noting that the positive/negative ef-

TABLE 2

MEAN RATINGS FOR PROFESSOR SKILL, INTERACTION, AND EXCELLENCE AS A FUNCTION OF POSITIVE/NEGATIVE TEACHING EVALUATIONS AND LECTURE PRESENTATION

Professor characteristic	Positive evaluation		Negative evaluation	
	Lecture 1	Lecture 2	Lecture 1	Lecture 2
Skill	4.48	3.93	4.36	3.64
Interaction	2.76	2.17	2.37	1.90
Excellence	4.48	3.46	3.74	3.08

fect held constant in both lecture conditions for all three dependent variables. Overall skill was rated as .41, interaction as .66, and excellence as 1.12 higher in the positive as compared to the negative evaluation condition.

Lecture presentation had a significant effect on the students' ratings of the professor's teaching skill ($F = 4.62$, $df = 1/184$, $p < .05$), his interaction capacity ($F = 17.86$, $df = 1/184$, $p < .01$), and his overall excellence ($F = 58.03$, $df = 1/184$, $p < .01$). Students in Lecture 1, as compared to students in Lecture 2, rated the professor as more skillful (4.42 versus 3.79, respectively), as interacting more with the students (2.57 versus 2.04, respectively), and as being a better professor overall (4.11 versus 3.27, respectively). Lecture 1 produced more favorable ratings than Lecture 2 for skill of the professor (1.27 higher), interaction capacity (1.06 higher), and overall excellence (1.68 higher).

Discussion

The experiment investigated the effects of preexisting positive/negative professor teaching evaluations and lecture presentation on ratings of teaching performance. An attempt was made to facilitate the generalizability to the results to the university environment by conducting part of the experiment under actual classroom conditions with regular students as subjects (Lecture 1). Furthermore, the stimulus materials used were evaluations that previously had been summarized and distributed to students by a university-wide, professor/course evaluation committee. It is likely that, had the experimenters constructed the evaluations, the stimulus materials may have had an even greater impact on the professor ratings. Lastly, the UMSU Teacher Rating Form was selected since it had been used previously to assess teaching ability at the University of Manitoba and since it is probably representative of the teacher rating forms that are presently being used in faculties and universities throughout the country.

Research on feedback effects in the college teaching environment has been minimal and existing data are somewhat equivocal (Pambookian, 1974). Typically, these studies have investigated the effect of teaching-evaluation feedback on future teaching behavior. However, the present experiment focused on a crucial methodological problem related to evaluation feedback: the impact on students' subsequent ratings of teaching performance. The results, consistent with findings in other areas of psychology, showed that existing expectations significantly influenced the ratings of the professor's teaching performance. Subjects who initially received the negative teaching evaluation rated the professor's lecture presentation lower in terms of the professor's capacity for interaction with students and of his overall excellence as a teacher in comparison to other university instructors. Skill of professor was not influenced by expectancy effects; however, the means were in the predicted direction. Of these three dimensions, the most important was considered to be overall excellence, and it was this dimension which showed the largest differences between the positive and negative groups.

These differences persisted in both lecture conditions, and there was an overall lecture effect. Although the positive/negative feedback effects existed in both lecture conditions, Lecture 2 showed a pronounced overall decline. The ratings of professor skill, interaction, and excellence exhibited a consistent drop for both positive (.55, .59, 1.02, respectively) and negative (.72, .47, .66, respectively) evaluation groups. Note that Table 2 indicates lecture material may interact with expectancy effects in that a "poor" teacher presenting interesting material is rated consistently higher in skill, interaction, and excellence than a "good" teacher presenting boring material. These data suggest that lecture content can be a major variable influencing performance ratings, and they give some credence to the argument of some professors who claim that by teaching certain courses like learning, statistics, and perception, their ratings suffer. Unfortunately, a confound exists in the experimental design in that the two lecture conditions differed in both content material

and testing situation, and consequently, it is difficult to determine precisely how these factors contributed to the final outcome.

One major criticism surrounding such techniques as the UMSU Teacher Rating Form concerns their validity for measuring "good" and "bad" teaching performance. Recently, both Leventhal (1973) and Knapper, McFarlance, and Scanlon (1972) have argued against earlier contentions that some teacher rating forms are valid (e.g., McKeachie, Lin, & Mann, 1971). In either case, the results of the present experiment suggest that the general distribution of professor ratings can have adverse consequences. If teacher rating forms are invalid, then their potential value to the professor as feedback to facilitate goal attainment must be questioned, especially in the case of negative evaluations. From a learning theory viewpoint, improper discriminative stimuli are being used as well as punishment for making the incorrect response. This is the least conducive and most inefficient method for achieving the desired response, effective teaching.

Assuming for the moment that teacher rating forms are both valid and reliable, consider the instructor who gets negative ratings similar to Professor A's in this study. Such ratings might have a direct impact on his teaching effectiveness in a variety of ways. First, the students who sign up for his course may be of poorer quality than the normal population. Those students who are academically oriented and who are concerned about learning and teaching may actively avoid those sections having negative evaluations. Second, the motivation level of the students may be similarly affected, with those having higher levels seeking those sections having positive evaluations. Third, we should expect the attitudes and expectations of those students enrolled in courses receiving negative evaluations to be changed in the direction of the ratings.

These developments, taken together, can reduce the teaching effectiveness of the professor since his course has selectively sampled an atypical group from the general population. In studying those factors influencing course selection, Leventhal, Breen, and Perry (1974) found that as many as 35% of 2,500 introductory psychology students indicated that a professor's teaching reputation was the primary determinant in their course selections. Obviously, the publicized professor/course evaluations exert a major influence on the professor's overall teaching reputation. The resultant effect is likely to produce a vicious circle in which negative evaluations cause poorer quality students to enroll in the course, creating adverse learning conditions, and eventually, negative reevaluations.

An argument can be made that this experiment was a single exposure study and, hence, is not truly representative of teaching conditions which extend over lengthy periods of time. Furthermore, a professor should be capable of improving his ratings, in time, by changing his teaching style. This seems a reasonable argument to make regarding expectancy effects, since the social psychological literature clearly shows that attitudes are modifiable. However, class selection procedures like those outlined above lead to nonrandom sampling biases in which the poorly rated professor ends up with poorer quality students. This creates adverse learning conditions which increase the probability that the professor will receive negative ratings again. Consequently, the publication of teaching evaluations may create a set of conditions which are extremely resistant to change, even over an extended period of time.

It is apparent that a professor evaluation system is needed in which the advantages of effective teaching are maintained, while the disadvantages of evaluation as demonstrated in this study are minimized. Within such a system, only the most *positive* evaluations would be publicized in the anticalendar, while average and poor ratings would be communicated to the respective professors for their own information. This would serve three functions. First, according to learning theory, it would reward effective teaching directly and increase its value through social recognition. Second, it would reduce the punishing effects of negative evaluations at least at the social level, with the result that the poorly rated profes-

sor would not be punished for not making the desired response. Third, for those professors who received poor ratings, this system would minimize biased selection procedures so that they do not receive an unrepresentative sample of poor quality students. In addition, an opportunity should be made available for professors to contest evaluations before they are made public. The professors should have access to the original data and to the machinery that was used to generate the written descriptions of the professor's teaching performance.

REFERENCES

Adair, J. G. *The human subject: The social psychology of the psychological experiment.* Toronto, Canada: Little, Brown, 1973.

Asch, S. E. Forming impressions of personality. *Journal of Abnormal and Social Psychology,* 1946, **41,** 258–290.

Isaacson, R. K., McKeachie, W. J., & Milholland, J. E. Correlations of teacher personality variables and student ratings. *Journal of Educational Psychology,* 1963, **54,** 110–117.

Kelley, H. H. The warm-cold variable in first impressions of persons. *Journal of Personality,* 1950, **18,** 431–439.

Knapper, C., McFarlance, B., & Scanlon, J. Student evaluation—An aspect of teaching effectiveness. *CAUT Bulletin,* 1972, **21,** 26–34.

Leventhal, L. Teacher rating forms: Critique and reformulation of previous validation designs. Paper presented at the College Teaching Symposium at the Annual Canadian Psychological Association Convention, Victoria, June 1973.

Leventhal, L., Breen, L., & Perry, R. P. The college teaching programme at the University of Manitoba. Presented at the Canadian Psychological Association Annual Convention, Windsor, Ontario, June 1974.

McKeachie, W. J., Lin, Y., & Mann, W. Student ratings of teacher effectiveness: Validity studies. *American Education Research Journal,* 1971, **8,** 435–445.

Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist,* 1962, **17,** 776–783.

Pambookian, H. S. Initial level of student evaluation of instruction as a source of influence on instructor change after feedback. *Journal of Educational Psychology,* 1974, **66,** 52–56.

Rosenthal, R. *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts, 1966.

Rosenthal, R., & Jacobson, L. Self-fulfilling prophecies in the classroom: Teacher's expectations as unintended determinants of pupils' intellectual competence. In M. Deutsch, I. Katz, & A. R. Jensen (Eds.), *Social class, race, and psychological development.* New York: Holt, Rinehart & Winston, 1968.

Skinner, B. F. *Verbal behavior.* New York: Appleton-Century-Crofts, 1957.

(Received November 13, 1973)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 857-863

COGNITIVE RESULTS BASED ON DIFFERENT AGES OF ENTRY TO SCHOOL: A COMPARATIVE STUDY

GILBERT R. AUSTIN[1]

Bureau of Educational Research and Field Service, University of Maryland

T. NEVILLE POSTLETHWAITE

International Institute for Educational Planning, Paris, France

This review is a secondary analysis of work done by the International Association for the Evaluation of Educational Achievement. It concerns mathematics, reading comprehension, and science. The objective was to look for a relationship between age of entry into school and achievement as measured at ages 10 and 13. Principal findings suggest a specificity of effect between early entrance into school and greater achievement in mathematics but not in reading comprehension or science. The authors suggest that gain in achievement based on early education may be due to the sensory motor experiences which are so commonly part of early childhood education programs. It is further suggested that the carefully planned and sequenced curriculum approach to teaching mathematics may have an effect here.

Interest in early childhood education has grown rapidly in Europe in the last few years, reflecting an increasing demand for equality of educational opportunities, the changing role of the family in society, and widening opportunities outside the home for women. Concurrent with the redefinition of family roles is a significant shift from education of young children as the primary responsibility of parents to a shared responsibility with institutions of education.

In the last 13 years, much has been written about the importance of early childhood education as a basis for later school learning. Hunt (1961) and Bloom (1964) are important contributors to this literature. An outgrowth of these authors' writings was the conception and development of such projects in the United States as "Head Start" and "Follow Through." The American efforts have been duplicated in a number of countries including England, with the Educational Priority Act; the Netherlands, with early compensatory education programs; and the Federal Republic of Germany, with a wide variety of cognitively oriented preschool programs. Many of these projects, in both North America and Europe, have made the assumption that early entry into school would have a beneficial effect on the children's later school achievement. In a number of national studies in Canada, the United States, Sweden, and the Netherlands, the benefits of these early school intervention programs, particularly any long-term effects, have been very difficult to document (Bissel, 1973; Davie, Butler, & Goldstein, 1972; Halsey, 1972; Kohnstamm, 1968; Plowden Report, 1967; Ryan, 1971; Stukat, 1971).

Method

The studies identified above have primarily been national efforts whose primary or secondary purpose is to address the question of effective age of entry into school. Since very little variation in effective age of entry into school occurs within a country, these studies are of limited value. An alternate method of looking at this question is to examine the results of multinational studies whose primary goal is not to answer questions about age of entry into school but whose data banks can be used in secondary analyses to give some insights about the problem.

[1] Requests for reprints should be sent to Gilbert R. Austin, Bureau of Educational Research and Field Services, College of Education, University of Maryland, College Park, Maryland 20742.

It is becoming more common to undertake secondary analyses since in large data collections the temporal and financial constraints on such projects usually mean that only a fraction of the possible analyses have been undertaken. This article has taken data from three International Educational Achievement (IEA) studies—mathematics (Husen et al., 1967; Postlethwaite, 1967), science (Comber & Keeves, 1973–1974) and reading comprehension (Thorndike, 1973–1974)—reanalyzed them in relationship to both the official age of entry and effective age of entry from Organization for Economic Cooperation and Development studies (Austin, 1973a, 1973b).

To discover whether the age at which students enter school is causally related to some specified later achievement, it would be necessary to assign students at random to various groups and have each group start at a different age but be subjected to identical schooling. While theoretically sound, such an experimental research would be impossible to conduct in practice within any one educational system.

The mean scores can be found in the original publications. Each country gave its *official* age of entry. Austin recategorized countries by *effective* age of entry, the year in which 75% or more of an age group enter formal school. The IEA studies are cross-sectional surveys with small standard errors of sampling.

Scattergrams were produced showing the relationship between both official age of entry, effective age of entry and mean scores.

TABLE 1

Official and Effective Age of Entry, Mean Scores, and Standard Deviations in Mathematics (Age 13)

Country	OAE (years)	EAE (years)	Population 1A		Population 1B	
			M	*SD*	*M*	*SD*
Australia	6	5	20.2	14.0	18.9	12.3
Belgium	6	3	27.7	15.0	30.4	13.7
England	5	5	19.3	17.0	23.8	18.5
Finland	7	7	15.4	10.8	16.1	11.6
France	6	4	18.3	12.4	21.0	13.2
Germany (Federal Republic)	6	6			25.5	11.7
Israel	6	4			32.2	14.7
Japan	6	6	31.2	16.9	31.2	16.9
Netherlands	6	4	23.9	15.9	21.4	12.1
Scotland	5	5	19.1	14.6	22.3	15.7
Sweden	7	7	15.7	10.8	15.3	10.8
United States	6	5	16.2	13.3	17.8	13.3

Note. Abbreviations: OAE = official age of entry, EAE = effective age of entry.

TABLE 2

Official and Effective Age of Entry, Mean Scores, Standard Deviations in Reading Comprehension and Science Age 10)

Country	OAE (years)	EAE (years)	Reading comprehension		Science	
			M	*SD*	*M*	*SD*
Belgium (Flemish)	6	3	17.5	9.2	17.9	7.2
Belgium (French)	6	3	17.9	9.3	13.9	7.1
England	5	5	18.5	11.6	15.7	8.5
Finland	7	7	19.4	10.8	17.5	8.2
Germany (Federal Republic)	6	6	—	—	14.9	7.4
Hungary	6	5	14.0	9.8	16.7	8.0
Israel	6	4	13.9	11.0	—	—
Italy	6	5	21.6	9.6	17.5	9.1
Japan	6	6	—	—	21.7	7.6
Netherlands	6	4	17.7	9.5	15.3	7.6
Scotland	5	5	18.4	11.1	14.0	8.3
Sweden	7	7	21.5	10.5	18.3	7.3
United States	6	5	16.8	11.6	17.7	9.3

Note. Abbreviations: OAE = official age of entry, EAE = effective age of entry.

Results

Tables 1 through 3 and Figures 1 through 4 present the data for national mean scores for 13-year-olds for mathematics (Populations 1a–1b), 10-year-olds and 14-year-olds for reading comprehension, and 10-year-olds for science and their relationships to official/effective ages of entry.[2]

It will be seen that earlier age of entry is associated with higher scores in mathematics at age 13 (Figures 1 & 2) and with lower scores in reading and science at age 10 (Figures 3 &4).

Discussion

Other data collected by IEA (Comber & Keeves, 1973–1974) indicate that science is spasmodically taught up to the age of 10 in most schools. Thus, it may not be surprising that the longer children remain at home

[2] Detailed definitions of populations 1A and 1B may be found in the original *International Study of Achievement in Mathematics* (Husen, 1967) document.

TABLE 3

OFFICIAL AND EFFECTIVE AGE OF ENTRY, MEAN SCORES, STANDARD DEVIATIONS IN READING COMPREHENSION (AGE 14)

	OAE (years)	EAE (years)	Reading comprehension	
			M	*SD*
Belgium (Flemish)	6	3	24.6	9.6
Belgium (French)	6	3	27.2	8.7
England	5	5	25.3	11.9
Finland	7	7	27.1	10.9
Germany (Federal Republic)	6	6		
Hungary	6	5	25.5	9.8
Israel	6	4	22.6	12.8
Italy	6	5	24.0	9.2
Japan	6	6		
Netherlands	6	4	25.2	10.1
Scotland	5	5	27.0	11.5
Sweden	7	7	25.6	10.8
United States	6	5	27.3	11.6

Note. Abbreviations: OAE = official age of entry, EAE = effective age of entry.

the higher their achievement on the IEA science tests.

In the IEA cross-sectional surveys, there was variability between countries and particularly within countries on the age of entry to school. The number of observations (in this case, national mean scores) are limited, typically to 12 or 13. Clearly the number of factors associated with between-country differences are manifold. Therefore, to examine only the relationship of national age of entry to later achievement is a highly speculative venture. Nevertheless, the contradictory results between reading and mathematics are so unexpected that they are worth presenting.

These differences in mathematics and reading comprehension achievement are not simply explained. Many factors influence the mean level of performance and the variation from the mean. *One* possible source of variation is, as we have indicated, age of entry, official or effective. It might be that mathematics instruction reflects the effects of early school intervention more than does reading comprehension. Such an assertion is based on the development sequence of children proposed by Piaget and Montessori, both of whom indicate that sensorimotor learning preceded symbolic or abstract learning. It is possible that the sensorimotor-play activities of preschool transfer to beginning mathematics skills more easily

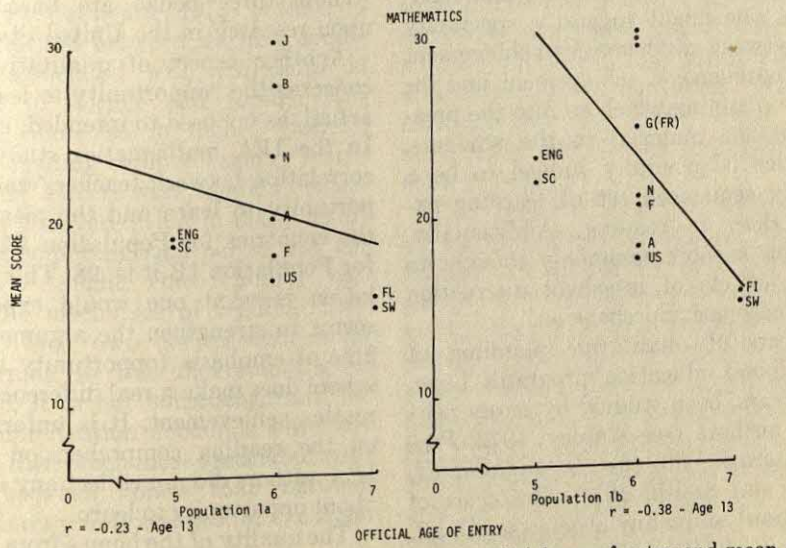


FIGURE 1. Scattergram showing relationship between official age of entry and mean scores for mathematics. (Abbreviations: J = Japan, B = Belgium, N = Netherlands, A = Australia, F = France, US = United States, Eng = England, Sc = Scotland, Fl = Finland, Sw = Sweden, G (FR) = Germany (Federal Republic), and Is = Israel.)

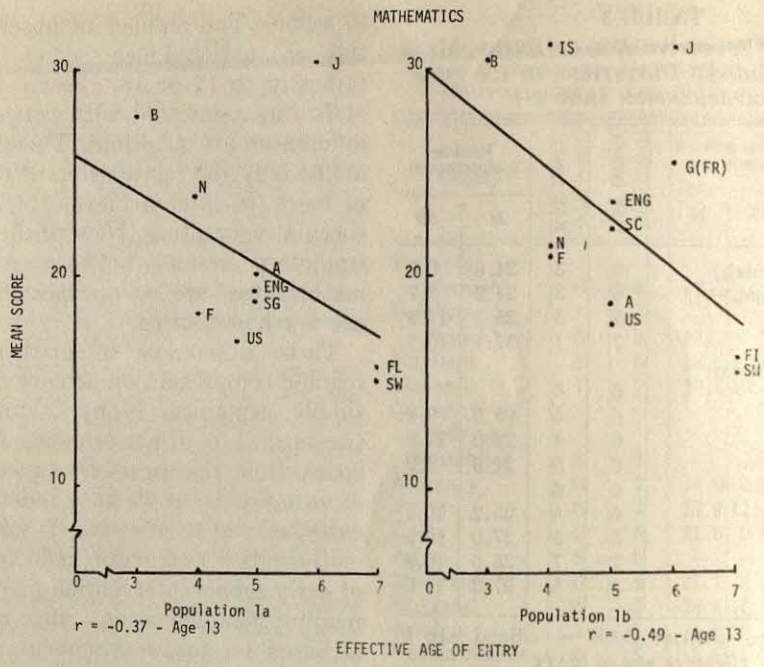


FIGURE 2. Scattergram showing relationship between effective age of entry and mean scores for mathematics. (Abbreviations: J = Japan, B = Belgium, N = Netherlands, A = Australia, F = France, US = United States, Eng = England, Sc = Scotland, Fl = Finland, Sw = Sweden, G (FR) = Germany (Federal Republic), and Is = Israel.)

than they do to the acquisition of reading comprehension skills.

In examining the mathematics and reading results, one might suggest a specificity of effect between mathematics achievement and/or comprehension achievement and the qualitative planning which go into the presentation of the material to the students. Mathematics is generally judged to be a more easily sequenced set of learning experiences than is reading. Additionally, mathematics is more commonly thought to reflect the effects of in-school instruction than does reading comprehension.

The effects of qualitative planning of early childhood education programs have, in recent years, been studied by many people. These authors (see Stanley, 1972, *Pre-School Programs for the Disadvantaged*, and Little and Smith, 1971, *Strategies of Compensation*) and many others (see Chall, 1967, *Learning to Read: The Great Debate*) conclude that to bring about qualitative changes in early childhood education, it is necessary to plan carefully for the attainment of stated goals and objectives and to evaluate the success of their attainment. (These three books are based primarily upon research in the United States).

Another aspect of qualitative planning concerns the "opportunity to learn" (or the actual, as opposed to intended, curriculum). In the IEA mathematics study, the rank correlation between teachers' ratings of opportunity to learn and the mean score for the countries for Population 1A is .96 and for Population 1B it is .98. This high correlation is what one would expect, and it seems to strengthen the argument that degree of emphasis (opportunity to learn) in school does make a real difference in mathematics achievement. It is unfortunate that on the reading comprehension study, the IEA authors did not collect any information about opportunity to learn.

The quality of the homes from which children come might also have an influence, although it may vary somewhat from country to country. The curriculum (the extent to which the educational objectives embodied

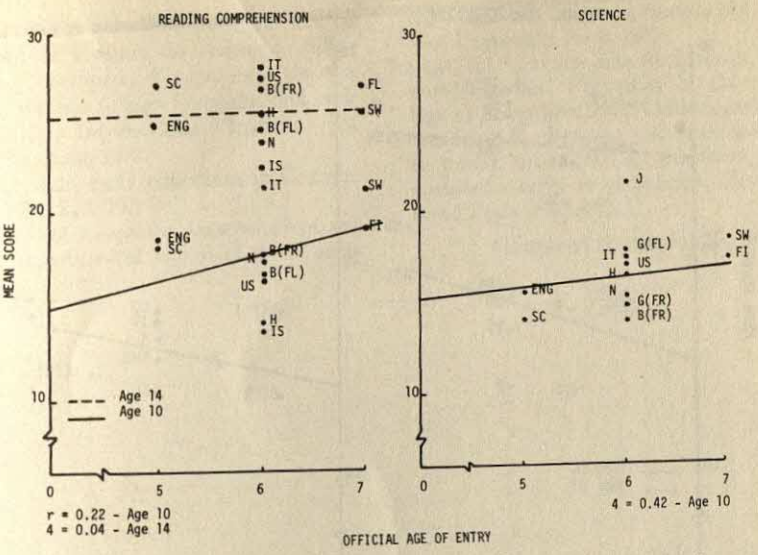


FIGURE 3. Scattergram showing relationship between official age of entry and mean scores for reading and science. (Abbreviations: IT = Italy, US = United States, SC = Scotland, B (FR) = Belgium (French), FL = Finland, ENG = England, H = Hungary, B (FL) = Belgium (Flemish), N = Netherlands, IS = Israel, SW = Sweden, J = Japan, and G (FR) = Germany (Federal Republic).)

in the materials used in school stretch the children) as well as the amount of homework will be important (cf. Postlethwaite, 1973).

One of the major findings of the total IEA study (Postlethwaite, 1973) is that, on average, variations between learning conditions were associated with variations in achievement in science and foreign languages (and more so in higher grades), whereas for reading comprehension "learning conditions" effects were small. Furthermore, it was discovered that in many countries reading comprehension per se was not systematically taught. Once children had conquered the mechanics of reading, they were left to their own devices concerning further learning in the understanding of texts. Hence, it is not surprising that the reading comprehension rebounds onto the home and is then much more associated with variations between homes than between schools. This may well account for the slope of 10-year-old scores in Figure 4.

The above interpretation is speculative, even though other factors have been controlled. The interpretation of the "effect" of age of entry is even more speculative, given no control of other factors. The problems of how to control for other factors in between-country comparisons have yet to be tackled.

This review suggests that there may be some interaction between age of entry, sensorimotor activities, qualitative planning, and the effects of home background on the attainment of instructional objectives in mathematics. The same variables should be investigated for reading comprehension achievement.

This secondary analysis exemplifies many of the difficulties inherent in such undertakings. To attack the specific problem of the effect of age of entry, a special experimental design study should have been conducted. Hence, the IEA design is inappropriate. Second, the lack of degrees of freedom in the number of observations is a problem. These, and similar problems, will always face researchers using data banks created for purposes different than their own. Nevertheless, in view of time and financial constraints properly executed, secondary analyses of primary data are a valid and valuable method of speculative analysis such as that presented in this study.

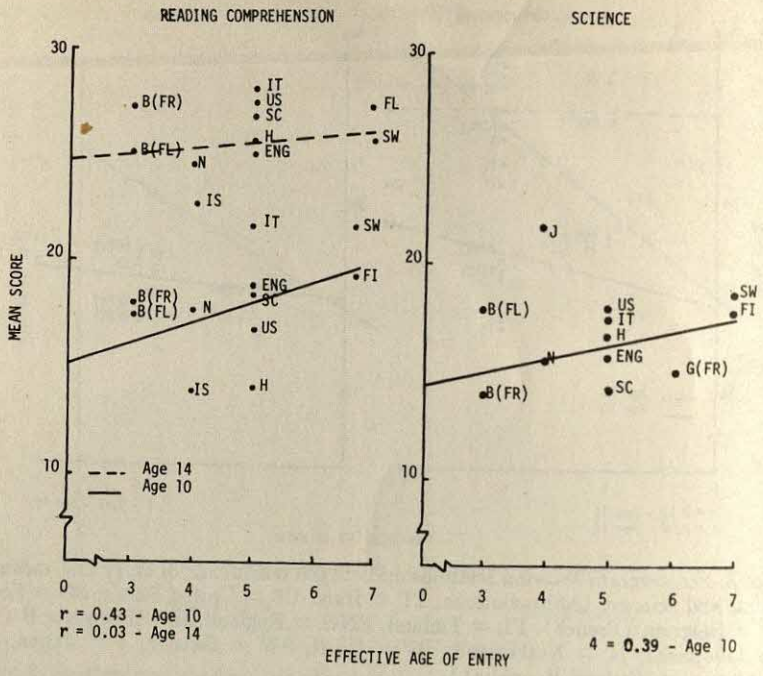


FIGURE 4. Scattergram showing relationship between effective age of entry and mean scores for reading comprehension and science. (Abbreviations: IT = Italy, US = United States, SC = Scotland, B (FR) = Belgium (French), FL = Finland, ENG = England, H = Hungary, B (FL) = Belgium (Flemish), N = Netherlands, IS = Israel, SW = Sweden, J = Japan, and G (FR) = Germany (Federal Republic).)

REFERENCES

Austin, G. Early childhood education in eight countries. Paris: Organization for Economic Co-operation and Development, 1973. (Mimeograph) (a)

Austin, G. International comparative study of school-related achievement as it reflects effects of early childhood education. Paris: Organization for Economic Cooperation and Development, 1973. (Mimeograph) (b)

Bissell, J. Planned variation in Head Start and Follow Through. In Julian C. Stanley (Ed.), *Compensatory education for children ages 2 through 8, recent studies of educational intervention.* Baltimore, Md.: Johns Hopkins University Press, 1973.

Bloom, B. *Stability and change in human characteristics.* New York: Wiley, 1964.

Chall, J. *Learning to read: The great debate.* New York: McGraw-Hill, 1967.

Comber, L., & Keeves, J. *Science education in nineteen countries: An empirical study. International studies in evaluation.* New York: Halsted Press, 1973–1974.

Davie, R., Butler, N., & Goldstein, H., *From birth to seven: A report of the national child development study.* London: Longman Group, 1972.

Halsey, A. *Educational priority: E.P.A. problems and policies.* Vol. 1. London: Her Majesty's Stationery Office, October 1972.

Hunt, J. *Intelligence and experience.* New York: Ronald Press, 1961.

Husen, T. et al. *International project for the evaluation of educational achievement. Phase I: International study of achievement in mathematics: A comparison of twelve countries.* Vols. 1 & 2. New York: Wiley, 1967.

Kohnstamm, G. *Project compensation programme.* (Project Report) the Netherlands: University of Utrecht, Department of Education, December 12, 1968. (Mimeograph)

Little, A., & Smith, G., *Strategies of compensation: A review of educational projects for the disadvantaged in the United States.* Paris: Organization for Economic Cooperation and Development, 1971.

Plowden. *Children and their primary schools* (The Plowden Report). Vols. 1 & 2. Central Advisory Council for Education. London: Her Majesty's Stationery Office, 1967.

Postlethwaite, T. N. *School organization and student achievement: A study based on achievement in mathematics in twelve countries.* Stockholm: Almqvist & Wiksell, 1967.

Postlethwaite, T. N. *A selection from the overall findings of the IEA study in science, reading comprehension, literature, French as a foreign language, English as a foreign language, and civic education.* Paris: International Institute for Educational Planning, 1973.

Ryan, T. Poverty and early education in Canada. *Interchange,* 1971, **2,** 1–10.

Stanley, J. *Preschool programs for the disadvantaged: Five experimental approaches to early childhood education.* Baltimore, Md.: Johns Hopkins University Press, 1972.

Stukat, K. *Forskolan och Skolstarten.* (Project Report) Sweden: University of Gothenburg, College of Education, 1971. (Mimeograph)

Thorndike, R. *Reading comprehension education in fifteen countries: An empirical study. International studies in evaluation.* New York: Halstead Press, 1973–1974.

(Received November 20, 1973)

Journal of Educational Psychology
1964, Vol. 55, No. 3, 864-871

LESSON KINETIC STRUCTURE ANALYSIS AS RELATED TO PUPIL AWARENESS AND ACHIEVEMENT

ROBERT J. BROWNE[1]

Hunterdon County Department of Education, Flemington, New Jersey

O. ROGER ANDERSON

Teachers College, Columbia University

This study investigated the effects of variations in communication structure on knowledge acquisition and subject's awareness of structure variation. Three transcripts differing in content structure as determined by kinetic structure analysis were developed. The amount of kinetic structure in a communication was the number of words held in common in pairs of contiguous statements. The 191 subjects were ninth graders. Three experimental groups were administered one of the three structure treatments. Data showed that student achievement is directly related to communication structure ($p < .05$). Subjects were not able to detect variations in lesson structure. A method of sensitivity training was devised which produced discrimination ability ($p < .05$). The theory-based, quantitative method of communication structure used herein yields significant effects on learning outcomes. Pupil assessment of teacher competencies in this domain may not be valid without preliminary sensitivity training.

Verbal communication accounts for a large proportion of classroom learning activity. Clearly the kind of content communicated and particularly the way it is organized can influence student knowledge acquisition and possibly affect student motivation to maintain sustained interest in the learning activity. The authors have performed a study of teacher verbal communication to determine the effects of variation in communication sequential organization on content acquisition and to examine the students' ability to assess variations in communication organization. This latter point is becoming of greater importance in the light of the movement toward student participation in teacher accountability models. This study investigated the basis for untrained observers, namely, high school students, participating in the evaluation of classroom teachers.

Previous research (Kissler & Lloyd, 1973; Natkin & Moore, 1972) on the effects of sequential organization on learning meaningful material has shown that criterion effects were observed most frequently when the verbal material contained interrelated sentences as opposed to free-standing sentences. A limitation of these studies, however, was that they did not provide a theoretical basis for variations in sequential organization nor did they use quantitative methods to assess variations in structure. In this study, we used a content analysis system devised by Anderson (1970, 1971) to generate various amounts of organization in curriculum content while holding nearly constant the total amount of content presented. This system provides a quantitative method to assess the amount of organization in a verbal communication and was used to verify that criterion levels of structure in the stimulus content matter had been realized. We therefore gained careful control over variations in stimulus organization, which was the independent variable in our study. A summary of the theory underlying this method is presented in background for a more specific statement of the research problem.

Theory

A basic measure of the amount of structure in a communication is the extent to

[1] Requests for reprints should be sent to Robert J. Browne, Hunterdon County Department of Education, County Administration Building, Flemington, New Jersey 08822.

which individual utterances are connected through linking ideas. If we consider a verbal communication to be composed of a series of statements approximately equivalent to a sentence in written discourse, then the amount of structure in the communication is defined in part as the amount of repeated substantive terms in pairs of contiguous statements in the communication. One or more repeated substantive terms (words) in contiguous statements produces linkage or continuity of thought in that part of the communication sequence.

For purpose of clarity and precision we call a single verbal utterance a discourse unit, and each of the substantive words (nouns, adjectives, and verbs representing objects, classes of things, or actions) within the discourse unit is called a verbal element. To assess the amount of structure in a communication, we examine each successive pair of discourse units in the communication and determine how many of the verbal elements in each pair are matched. From this, we obtain a coefficient of content organization called the fundamental coefficient. It is a measure of the amount of matching of verbal elements in consecutive pairs of discourse units. The concept of matching, or of linking of ideas as it were, between consecutive discourse units is called commonality. In Figure 1, the concept of commonality is illustrated for three consecutive verbal statements represented as rectangles containing words represented by alphabetic letters. The arrows between consecutive rectangles show the commonality in each pair. The fundamental coefficient (B_1) is a measure of commonality.

$$B_1 = \frac{n_1}{n_0 + n_1},$$

where n_1 is the total number of matched verbal elements in a discourse unit pair, and n_0 is the number of unmatched verbal elements. A B_1 coefficient is computed for each discourse-unit pair. For example, in Figure 1, the first pair of discourse units contains a total of four matched verbal elements. There are two unmatched elements. Therefore, $B_1 = 4/2 + 4 = .67$. The second successive pair has two matched elements and three unmatched elements; thus $B_1 = 2/3 + 2 = .40$. When analyzing a transcript each verbal element is assigned a code number and is represented by that number each time it appears in successive discourse units.

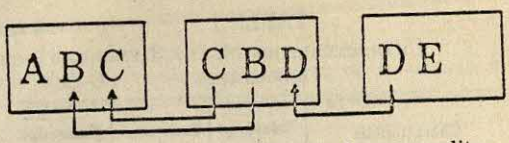


FIGURE 1. A conceptual schema of commonality.

In addition to B_1, a second coefficient, B_2, was produced. The coefficient B_2 includes a potency factor, a ratio whose magnitude depends on the frequency of occurrence in the entire lesson of the unmatched elements in a discourse-unit pair. The B_2 yields a graphical analysis of the transcripts called kinetograms, as explained in a previous publication (Anderson, 1971).

In addition to commonality as a variable in our research, we define a theme as being a linking idea that occurs repetitively throughout a section of the communication. The amount of theme activity is the frequency of occurrence of the linking idea divided by the total number of discourse units in that section containing the theme.

Purposes

The purposes of this study were to (*a*) systematically vary the amount of commonality in a verbal communication; (*b*) determine the effects of this variation on acquisition of the communicated knowledge; and (*c*) determine whether students were aware of the variations in content organization. There is currently much interest in teacher accountability and in student evaluation of teacher performance. A clear need exists to obtain evidence about the kinds of teacher behaviors that actually influence student learning and to determine under what conditions, if at all, students can assess variations in relevant teacher performance.

Hypotheses

The conceptual hypotheses of this study were the following: (*a*) Student content achievement is directly related to lesson content structure; and (*b*) awareness of commonality and theme development is directly related to lesson content structure.

TABLE 1
CHARACTERISTICS OF THE TAPED TRANSCRIPTS

Characteristic	Transcript 1	Transcript 2	Transcript 3
B_1	.54	.41	.22
Discourse units	88	88	89
Coded verbal elements	74	74	75
Verbal element frequency	427	428	355
Elapsed time of tape (min.)	9.75	8.83	7.07
Discourse units per minute	10	11	13

METHOD

Preparation of Treatment Communications

The characteristics of the three transcripts used to prepare tape-recorded communications are reported in Table 1. Discourse units per minute were computed by averaging the number of discourse units spoken in the corresponding minutes of the three transcripts. Sample Minutes 1, 2, and 5 were chosen randomly with the aid of a table of random numbers. While the recorded transcript was played, the discourse units in the first, second, and fifth minutes were counted and totaled. The total was divided by three to obtain the average number of discourse units spoken per minute. This process was repeated for each of the three transcripts.

The high-kinetic-structure transcript, mean B_1 of .54 was produced by making minor revisions to the transcript used by Trindade-Khristanand (1971).

The intermediate-kinetic-structure transcript used in this study was generated from the high-kinetic-structure transcript by reorganizing the discourse units. An attempt was made to retain content logic—to have the script make sense when read by preserving the logical meaning of connecting words such as "therefore" between sentences. Repetitious discourse units and statements giving examples were removed and associated with different discourse units, sometimes as short reviews. Care was taken not to have terms used before they were defined. The mean B_1 of the intermediate-kinetic-structure transcript was .41. The low-kinetic-structure transcript was generated by reorganizing the intermediate transcript. It had a mean B_1 of .22. Samples of the coded transcripts and verbal element code list of the three transcripts used in this study are reported in Table 2. The difference in the amount of commonality, as measured by mean B_1, is shown in Table 1.

Each of the transcripts was analyzed by plotting kinetograms to determine how consistent the level of structure was throughout the lessons. Based on this data we found each of the lessons of uniform structure throughout.

To aid the students with the conceptualization of the scientific terms used in the transcripts, a diagram was provided. It consisted of a pen and ink drawing of the major structures of *Rhizopus*. Through conferences held with the elementary school science teachers it was ascertained that the subjects would be naive with reference to the content of the treatment sessions, namely, the structure and life cycle of the bread mold, *Rhizopus*. With this knowledge it was decided not to use a pretest and thus to reduce subject fatigue.

Preparation and Scoring of Evaluation Instruments

Knowledge acquisition was measured by a 30-item, four-part, multiple-choice content achievement test. The test questions were designed to assess the subjects' knowledge of the structure and life cycle of *Rhizopus*. Trindade-Khristanand used the same test in his study and reported a test reliability of .89 (Trindade-Khristanand, 1971).

A questionnaire was prepared to assess pupil awareness of commonality. Statements selected for the final version had a Likert-type scale of one through seven. Statement A read: "In this lesson the same idea is repeated when moving from one statement to the next"; and Statement B read: "Each idea presented is closely related to the preceding idea."

The scales of the awareness questionnaire were scored according to where the subject recorded an *X*. Intervals along the scale were designated by indices every half point from 1.0 through 7.0. A mark anywhere in the interval between 3.0 and 3.5, for example, was scored as 3.0.

Treatment Groups

Subjects were randomly assigned to treatment groups. The mean IQ of the 191 students who participated in the study was 105.82. An analysis of variance showed no significant difference among the means of the three treatment groups, with respect to IQ. Treatment group sessions were held in the high school. The treatment session followed a standard protocol, namely, reading to the students the directions for the study and then playing the audiotape with the *Rhizopus* diagram in full view of the subjects. Next, the content achievement test was administered and then the awareness questionnaire.

Structure Awareness Training Methodology

As reported in the Results section, the findings from the full-scale study on the questionnaire showed that the subjects failed to discriminate among the levels of structure. A collateral study was carried out to investigate the conditions necessary for subjects to make this discrimination. Several classes of biology students from the same

TABLE 2

SAMPLES OF CODED TRANSCRIPTS

Discourse unit	Script	Verbal element
	High-kinetic-structure transcript	
1	Today we are going to study the history or life cycle of the bread mold, *Rhizopus*.	1, 2
2	Bread mold, whose history we're going to investigate, is a plant.	2, 1, 3
3	We come to know a plant by carefully noting down its external appearance and characteristics, as here in the case of bread mold.	3, 4, 2
4	We will study how a plant grows, particularly how bread mold matures and reproduces itself.	3, 5, 2, 6, 7
5	There are two ways in which a plant reproduces itself: (*a*) sexually and (*b*) asexually.	3, 7, 8, 9
6	The plant whose reproduction, both sexual and asexual, we are going to consider is a fungus called *Rhizopus* or bread mold.	3, 7, 8, 9, 10, 2
7	We will study its external appearance and its reproduction, sexual and asexual.	4, 7, 8, 9
	Intermediate-kinetic structure transcript	
1	Today we are going to study the history or life cycle of the bread mold, *Rhizopus*.	1, 2
2	We come to know a plant by carefully noting down its external appearance and characteristics, as here in the case of bread mold.	3, 4, 2
3	We will study how a plant grows, particularly how bread mold matures and reproduces itself.	3, 5, 2, 6, 7
4	Bread mold does not have chlorophyll and represents a certain type of plant in the plant kingdom.	2, 11, 3, 12
5	There are two ways in which a plant reproduces itself: (*a*) sexually and (*b*) asexually.	3, 7, 8, 9
6	The plant whose reproduction, both sexual and asexual, we are going to consider is a fungus called *Rhizopus*, or bread mold.	3, 7, 8, 9, 10, 2
7	We will study its external appearance and its reproduction, sexual and asexual.	4, 7, 8, 9
	Low-kinetic-structure transcript	
1	Today we are going to study the history or life cycle of bread mold, *Rhizopus*.	1, 2
2	We come to know a plant by carefully noting down its external appearance and characteristics.	3, 4
3	We study plants by investigating how a plant grows, matures, and reproduces itself.	3, 5, 6, 7
4	Bread mold is a plant.	2, 3
5	Reproduction is the formation of new individuals having characteristics typical of their family, thus bread molds reproduce bread molds as cats reproduce kittens.	7, 4, 64, 2, 65, 66
6	There are two types of reproduction: sexual and asexual.	7, 8, 9
7	Sexual reproduction is brought about by the union of two nuclei in male and female sex cells.	8, 7, 67, 47, 68, 69, 70

(*Note to Table 2 appears on following page.*)

TABLE 2—*Continued*

Note. Verbal element codes are as follows:

1 History, life cycle
2 Bread mold, *Rhizopus*
3 Plant, plant body
4 External appearance
5 Grow, growth, develop
6 Mature, ripe
7 Reproduce, reproduction
8 Sexual, sex
9 Asexual
10 Fungus, mold, rust
11 Chlorophyll
12 Plant kingdom
47 Nucleus
64 Family
65 Cat
66 Kitten
67 Union, unite, fuse
68 Male, plus strain
69 Female, minus strain
70 Sex cell

high school participated in a study of the following factors which could have caused the subjects to fail in discriminating among the levels of kinetic structure in the transcripts.

The instructions for the awareness questionnaire directed the subjects to compare the recorded transcript with all the lessons they have ever had in their experience. This might have been too general a comparison. In order to get an indication of the strength or weakness of this comparison, two biology classes were presented with the first half of the recording of the low-kinetic-structure transcript ($\overline{B}_1 = .22$) and the first half of the high-kinetic-structure transcript ($\overline{B}_1 = .54$). In one class, the low-kinetic-structure tape was played first and in the other class the high-kinetic-structure transcript was played first. The students were directed to compare the taped transcripts to one another by using the same questionnaire as in the full-scale study. The results are reported in Table 7. This direct comparison of the low- and high-kinetic-structure lessons did not enable them to discriminate the levels of structure.

Another factor that might interfere with the subjects' discrimination among the levels of kinetic structure is the content of the statements of the questionnaire. In order to gain more information concerning this factor, two more classes were presented with the comparative evaluation task as outlined above. They were asked to listen to the first half of the two transcripts and then use the questionnaire of the full-scale study to compare them. Again the position of the transcripts in the treatment was alternated. This time before playing the tapes the students were given a brief introduction that explained commonality as the repetition of words in statements next to one another.

TABLE 3

MEAN SCORES OF THE CONTENT ACHIEVEMENT TEST

Transcript	*N*	Mean test score	*SD*
High	65	14.26	3.568
Intermediate	63	13.71	3.270
Low	63	12.81	3.156

The students were told that the two lessons differed on this point and that after hearing the tapes they would be asked to compare the lessons. The results are reported in Table 7. With this brief awareness training the students were able to discriminate the lesson kinetic-structure levels.

The training procedure was repeated with two other classes but this time they only heard one tape, either the complete transcript of the low-kinetic-structure lesson or the complete transcript of the high-kinetic-structure lesson. They were then asked to use the questionnaire of the full-scale study to compare the transcript with all the other lessons they had ever experienced. The results are reported in Table 7. With respect to Statement A, students were even able to discriminate structure level by using the very general comparison of the full-scale study—all the lessons in their experience.

All of these treatments were as similar to those of the full-scale study as possible, including the display of the *Rhizopus* diagram. The students were assigned to the various biology classes at random by a computer after they had been grouped by IQ and past performance in science. The 123 tenth-graders who participated in this study had an IQ range of 90–130, as measured by the Lorge-Thorndike test in the ninth grade. They had not yet studied the structure and life cycle of *Rhizopus*.

Data Analysis

The data of the content achievement test were used to test the hypothesis that student content achievement is directly related to lesson kinetic structure.

The data of the awareness questionnaire were used to test the hypothesis that the awareness of commonality and theme development is directly related to lesson kinetic structure. The data were analyzed by performing a one-way analysis of variance.

The data of the structure awareness training groups were evaluated by the *t*-test method. The significance level for the statistical tests was $p \leq .05$.

RESULTS

The first hypothesis on knowledge acquisition was supported. The results of the

TABLE 4

ANALYSIS OF VARIANCE OF MEAN SCORES OF THE CONTENT ACHIEVEMENT TEST

Source of variation	*SS*	*df*	*MS*
Between groups	68.633	2	34.317
Within groups	2095.128	188	11.144
Total	2163.760	190	

Note. Between groups, $F = 3.079$, $F_{.05} = 3.05$; $df = 2/188$, *MS*.

analysis of variance are reported in Table 4. The mean scores on the content achievement test fall in the predicted order (see Table 3).

The second hypothesis predicted that the subjects would perceive the level of kinetic structure analysis of the transcript they heard by comparing its commonality to all the lessons in their experience. Tables 5 and 6 show the results of this part of the study. Since both of the F values are below the stated acceptable level of $p \leq .05$, the second hypothesis was rejected.

Table 7 contains the data of the structure awareness training study. The mean awareness value was computed by averaging the value assigned by the subjects to the two statements on the awareness questionnaire. The questionnaires were evaluated in the structure awareness training study in the same manner as was described for the full-scale study. Table 8 contains the results of the evaluation of the data of the structure awareness training study by use of the t-test method. It is evident from the means in Table 7 and the probability values in Table 8 that giving the subjects a direct comparison of the low- and high-kinetic-structure lessons did not enable them to discriminate the levels of structure. The subjects were only able to perceive structure levels after structure awareness training. For the most part, they were even able to discriminate structure level by using the very general questionnaire item containing a comparison of the treatment lesson with all other lessons in their experience.

TABLE 5

AWARENESS QUESTIONNAIRE DATA

Transcript	*N*	*M*	*SD*
Questionnaire Statement A			
High	65	4.52	1.872
Intermediate	63	4.16	1.879
Low	63	4.24	2.123
Questionnaire Statement B			
High	64	4.45	1.766
Intermediate	63	3.88	1.802
Low	63	4.51	1.466

TABLE 6

ANALYSIS OF VARIANCE OF AWARENESS QUESTIONNAIRE DATA

Source of variance	*SS*	*df*	*MS*
Questionnaire Statement A			
Between groups	4.718	2	2.359
Within groups	722.557	188	3.843
Total	727.275	190	
Questionnaire Statement B			
Between groups	15.054	2	7.527
Within groups	531.162	187	2.840
Total	546.216	189	

Note. For Statement A, between groups, $F = .614$; $F_{.05} = 3.05$; total $df = 2/188$, *MS*. For Statement B, between groups, $F = 2.650$; $F_{.05} = 3.05$; total $df = 2/187$, *MS*.

DISCUSSION

Teaching is a nontangible phenomenon existing as a series of events from which we deduce certain conceptual categories called teaching acts. The identification of structure in teaching is not as straightforward as in material structure analysis. Therefore, analysis of structure in teaching requires even more careful definition and precise empirical indicators to enhance realization of reliable and professionally significant findings. The concept of structure used herein is based on clearly observable events in teaching and therefore may not account for all of the subtleties in organization considered significant in teaching. However, the added precision gained by careful empirical

definitions should yield enhanced reproducibility and precise application of findings to daily classroom teaching practice. The hypotheses tested in this study concern very particular aspects of the teaching process. This particular focus, it is hoped, will provide a clearer understanding of the finding than would be possible with more abstract, less empirical structure variable definitions.

As reported in the results section, the first hypothesis was confirmed; namely, that student content achievement will be directly related to lesson kinetic structure. The prediction of the outcome of the first hypothesis was based on a theory developed from biology and psychology (Anderson, 1971). Biologically, the theory holds that the organization of the nervous system of higher organisms is such that the organism is maximally sensitive to periodic stimulus input and that it forms relationships among similar stimuli occurring in close temporal succession. Psychologically, the theory holds that the learning process is enhanced when verbal communications contain linked verbal elements in contiguous verbal utterances thus producing proactive facilitation, that is, the facilitation of subsequent acquisition through the presentation of verbal elements that cue the learner to anticipate repetition and hence linkage of ideas. Communications of high kinetic structure contain more repetition and hence linkage of contiguous verbal elements (commonality) than do communications of low kinetic structure. The results of this study show that the amount of knowledge acquisition was directly related to the amount of structure in the treatment lessons. One must realize, however, that too much commonality might cause a loss in acquisition of content because of boredom with excessive repetition. Conversely, if there are too few common verbal elements among continguous dis-

TABLE 7
DATA OF THE STRUCTURE AWARENESS TRAINING STUDY

Structure awareness treatment group and no.	*M*	*SD*	Variance
Questionnaire Statement A			
1. No awareness training; first half of low and high transcripts; low is second tape played.	4.33	1.71	2.91
2. No awareness training; first half of low and high transcripts; high is second tape played.	4.59	1.34	1.81
3. Awareness training; first half of low and high transcripts; low is second tape played.	3.25	1.54	2.37
4. Awareness training; first half of low and high transcripts; high is second tape played.	5.79	1.60	2.56
5. Awareness training; only complete tape of low transcript played.	4.06	1.84	3.38
6. Awareness training; only complete tape of high transcript played.	5.46	1.26	1.59
Questionnaire Statement B			
1. No awareness training; first half of low and high transcripts; low is second tape played.	4.35	1.75	3.06
2. No awareness training; first half of low and high transcripts; high is second tape played.	4.46	1.45	2.11
3. Awareness training; first half of low and high transcripts; low is second tape played.	3.22	1.76	3.10
4. Awareness training; first half of low and high transcripts; high is second tape played.	5.76	1.55	2.42
5. Awareness training; only complete tape of low transcript played.	4.44	1.43	2.05
6. Awareness training; only complete tape of high transcript played.	4.39	1.00	1.01

TABLE 8

T-TEST ANALYSIS OF THE DATA FOR THE STRUCTURE AWARENESS TRAINING STUDY

Structure awareness treatment group and no.	Questionnaire statement	t	df	$p \leq .05$
1. No awareness training; first half of low and high transcripts; low is second tape played.	A	.59	46	*ns*
2. No awareness training; first half of low and high transcripts; high is second tape played.	B	.24	45	*ns*
3. Awareness training; first half of low and high transcripts; low is second tape played.	A	4.85	35	Significant
4. Awareness training; first half of low and high transcripts; high is second tape played.	B	4.65	35	Significant
5. Awareness training; only complete tape of low transcript played.	A	2.52	36	Significant
6. Awareness training; only complete tape of high transcript played.	B	.10	36	*ns*

course units the pupils may fail to associate the units in a series or expend so much work in supplying the links that they will become fatigued and knowledge acquisition will also be depressed. Therefore, in an efficient communication there must be a reasonable balance between commonality and the progressive introduction of lesson content.

The second hypothesis of this study, that awareness of commonality is directly related to lesson kinetic structure, was rejected. However, the results of the structure awareness training study show that subjects can be taught to discriminate among levels of kinetic structure. The cursory training consisted of defining commonality according to the concepts represented in Figure 1 of this article. The findings from this part of the study extend our knowledge of kinetic structure analysis by providing some information about how lesson kinetic structure is viewed by the learner. Moreover, these findings shed some light on possible errors in student ratings of teacher performance—the students may not be sufficiently aware of the variables to assess them validly. Thus, even though the students suffer a learning deficit when structure is low; as shown in this study, they sometimes cannot identify the source of the instructional flaw without some training in the assessment of instructional procedures. Much further research is needed to determine what kind of assessment training and what maturation effects influence student assessment of significant teacher performance variables.

REFERENCES

Anderson, O. R. A comparative analysis of structure in teacher communicated science content. *Journal of Research in Science Teaching*, 1970, **7**, 227–244.

Anderson, O. R. *Quantitative analysis of structure in teaching*. Presentation of the theory and methods of kinetic structure analysis. New York: Teachers College Press, 1971.

Kissler, G. R., & Lloyd, Kenneth E. Effect of sentence interrelation and scrambling on the recall of factual information. *Journal of Educational Psychology*, 1973, **64**, 187–190.

Natkin, G. L., & Moore, J. W. The effects of instructional sequencing on learning from a single knowledge structure. *American Educational Research Journal*, 1972, **9**, 599–605.

Trindade-Khristanand, A. Structures in science teaching and learning outcomes. Unpublished doctoral dissertation, Teachers College, Columbia University, 1971.

(Received January 3, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 872-879

EFFECT OF TWO FREE-TIME REINFORCEMENT PROCEDURES ON ACADEMIC PERFORMANCE IN A CLASS OF BEHAVIOR PROBLEM CHILDREN[1]

DAVID MARHOLIN II,[2] ELIZABETH T. McINNIS,[3] AND TOM B. HEADS[4]

University of Illinois at Urbana-Champaign and Herman M. Adler Center, Champaign, Illinois

Three behavior problem children, ages 11-14, who had been placed in a state mental health facility, were subjected to baseline and two sequential reinforcement conditions (reading reinforcement and chance reinforcement). The results indicate that free time can function as a reinforcer producing increases in the percentage of items answered correctly. Further, the results indicate that improvements can be affected in other academic subjects when only reading accuracy is reinforced. Finally, it was apparent that neither a reading reinforced nor a chance reinforcement condition produced consistently differential results. Results imply that a procedure which reinforces the qualitative aspects of academic performance can be effective in increasing the accuracy of academic performance. Moreover, these improvements in academic behavior were achieved using free time, a reinforcer which is inexpensive and readily available to the teacher. It should also be noted that a single 20-minute free-time period was effectively used to produce changes in three academic areas.

Behavioral control techniques have been successful in increasing frequencies of task-oriented behaviors and decreasing incompatible, disruptive behaviors in a variety of classroom populations. Madsen, Becker, and Thomas (1968) showed that by ignoring inappropriate behavior and attending to appropriate behavior, the inappropriate behavior of three elementary school children was effectively diminished. Similar results were obtained in classes of first-grade behavior problem children (Hall, Lund, & Jackson, 1968; Ward & Baker, 1968), with institutionalized female offenders (Meichenbaum, Bowers, & Ross, 1968), and with a retarded child (Brown, Montgomery, & Barclay, 1969). A study involving 28 well-behaved children in a middle-primary public school demonstrated that disruptive behavior resulted when the teacher withdrew positive attention for appropriate classroom behavior and was eliminated when contingent positive approval was reinstated (Thomas, Becker, & Armstrong, 1968). An experiment with five hyperactive children in a special class demonstrated that visual orientation toward the teacher could be controlled by social reinforcement (Quay, Werry, McQueen, & Sprague, 1966).

Token reinforcement programs have effectively increased the frequency of task-oriented behavior in classrooms of behavior problem children (O'Leary & Becker, 1967; O'Leary, Becker, Evans, & Saudargas, 1969). Walker, Mattson, and Buckley (1969) increased the on-task behavior of fourth-, fifth-, and sixth-grade disruptive

[1] This research was supported in part by Extramural Grant 250-11-RD, Illinois State Department of Mental Health.

The authors wish to thank the staff of the Herman M. Adler Center, with special thanks to Joy Yonko, classroom teacher, for her assistance and participation.

[2] Requests for reprints should be sent to David Marholin II, who is now at Psychological Clinic, Children's Research Center, 51 Gerty Drive, Champaign, Illinois 61820.

[3] Now at Los Ninos Remedial Center, 3725 30th Street, San Diego, California 92104.

[4] Now at the Mansfield Training Center, Mansfield Depot, Connecticut.

children from a base rate of 39% to a rate of 90% at the conclusion of a token reinforcement program. Broden, Hall, Dunlap, and Clark (1970) demonstrated similar increases in on-task behavior from a baseline period (29%) to treatment conditions (74%) with seventh and eighth graders who displayed more than average amounts of disruptive behavior during baseline. In a series of single-subject manipulations, Becker, Madsen, Arnold, and Thomas (1967) found that a combination of ignoring deviant behavior and reinforcing incompatible appropriate behavior was most effective in classroom management. Mattos, Mattson, Walker, and Buckley (1969) reported that reinforcement of appropriate behavior and punishment of deviant behaviors with fourth-, fifth-, and sixth-grade behavior problem children was effective in producing behavioral change. Such studies have aptly demonstrated that attention, lack of attention, or punishment delivered contingently upon a response can be effective in altering the frequency of classroom behaviors.

Several studies have investigated changes in academic performance due to token reinforcement systems. A 14-year-old culturally deprived juvenile delinquent increased more than two grade levels in reading achievement with an intensive one-to-one token program over a four-and-a-half month period (Staats & Butterfield, 1965). Nolan, Kunzelmann, and Haring (1967) reported similar results with eight junior high school students with serious learning and behavior disorders. Using a token reinforcement system, Birnbrauer, Wolf, Kidder, and Tague (1965) were successful in decreasing the percentage of errors in 10 of 15 retarded subjects. Wolf, Giles, and Hall (1968) found significant increases in school grades of low-achieving fifth and sixth graders, as compared with a control group, when reinforcement was made contingent upon mastery of standard instructional materials. In one of the largest token reinforcement studies with emotionally disturbed children, Hewett, Taylor, and Artuso (1969) found increases in task-oriented behaviors and standard achievement test scores over those obtained by control groups. These studies demonstrate empirical support for the effectiveness of token reinforcement programs to increase academic performance as a function of contingent payoffs.

Most operant classroom studies, however, have concentrated on systematically investigating the effects of reinforcement contingencies on disruptive and task-oriented behaviors. Classroom management techniques are a necessary precursor to a setting which is conducive to increasing academic skills and, indeed, many children are referred for residential treatment because of social behavior problems in the school setting. However, the level of academic functioning of these children also is frequently found to be far below grade level. Thus, a program designed to accelerate academic behaviors incompatible with the referral problems is necessary. Establishing only appropriate social behavior (e.g., sitting at desk and holding a pencil), though necessary, is not regarded as sufficient, since these children cannot be returned to public schools unless academic as well as social deficits are remedied. Maximizing the probability of natural environmental reinforcers (Baer, Wolf, & Risely, 1968) requires that academic behavior, as well as social behavior, be established.

This study was designed to investigate two questions. First, is free time (as demonstrated by Osborne, 1969) a reinforcer, as indicated by its contingent effect upon accuracy of academic performance? Second, does reinforcement which is contingent upon reading performance produce an increase in other subject areas (mathematics and English) and if so, is the overall result of reinforcing one subject greater than the result produced when reinforcement is unpredictably contingent upon performance in any one of the three subject areas?

METHOD

Setting

The three subjects were residents in an in-patient behavior modification program for emotionally disturbed children from 6 to 14 years of age.

They all attended special education classes in a school operated by the institution. A system of token reinforcement using marks which could be exchanged for a variety of back-up reinforcers was used in three classrooms. All six children in a classroom of four males and two females were included in the study. Their ages ranged from 11 to 14 years. Five of the six children possessed a history of acting out behavior in public school classrooms prior to admission to the program. Their classroom problems typically included such behavior as getting out of their seats without permission, fighting, calling out, using verbal abuse, throwing objects, and refusing to work. The children were on the average three years below grade level as determined by norms for their chronological age.

All classes were held in a single classroom which contained six desks arranged in rows, facing the teacher's desk which was at the front of the room. Two one-way viewing mirrors in addition to a camera mounted at ceiling level in one corner of the room allowed the class to be observed at all times from several perspectives. The class was directed by a certified special education teacher who was assisted by one of two rotating aides. Because three of the children were discharged before completion of the study, the reported data are based upon those three children who completed all conditions.

Design

An A-B-C-A intrasubject analysis was used (Sidman, 1960). The four conditions were an initial baseline (A), reading reinforced (B), chance reinforcement (C), and a final baseline (A).

Initial Baseline (A)

The teacher maintained a record of the number of items attempted and the number of correct problems for each child during the reading, mathematics, and English periods. The final 20 minutes of the morning school session was designated as the free-time period. During this time, students had access to games, record players, arts and crafts materials, special projects, and social interaction with the teacher, teacher's aide, and other children. Individuals earned 20 minutes of free time at the conclusion of the morning classes if they were not timed out from the class during the three academic periods. Behaviors punished by time out were classified into seven categories ranging on a continuum from minor offenses (e.g., noncompliance, talking without raising hand, moderate physical or verbal abuse) to more serious ones (e.g., stealing, assault on adults, major destruction, extortion). Duration of time out was a function of the seriousness of the offense, with times ranging from 10 minutes for minor offenses to 50 minutes for the most serious offenses. The time-out room, a short distance from the classroom, was 4 × 7 feet in size and devoid of all furniture, manipulable objects, and social stimuli. One verbal prompt was given for minor offenses before the time-out consequence was invoked. During free time, time out was instituted only for the more deviant behaviors (e.g., aggression toward others, swearing, destruction of property). The use of time out was consistent through out all conditions of the study.

TABLE 1

MEDIAN ACCURACY SCORES FOR EACH SUBJECT DURING INITIAL BASELINE

Subject	Reading (%)	English (%)	Mathematics (%)
1	80	75	85
2	95	92	95
3	80	98	92

Note. The median accuracy scores served as the criterion levels required for each subject to earn free time.

Reading Reinforced (B)

The median accuracy score for each child was determined from the data collected during the initial baseline (see Table 1). Free time was earned by each child if during the morning reading period he attained or surpassed his median accuracy score. The teacher continued to collect the rate and accuracy data in mathematics and English, although no specific contingency was based on these data. The teacher told each child what his accuracy criterion was and gave him an example of how it was determined. Those children not earning free time were given a 20-minute study hall period.

Change Reinforcement (C)

At the conclusion of the morning academic periods and immediately preceding free time, each student blindly drew one of three cards on which the three morning subjects were written. The card selected by the child designated the morning academic subject (English, mathematics, reading) upon which free time was contingent. The child had to meet his accuracy criterion, which was determined from the initial baseline (see Table 1). The teacher told each child what his criterion was in each subject and explained to him how he would select the subject that would determine free time. If during the chance reinforcement condition the child was absent from any of the academic periods, he drew from the remaining subjects. As in previous conditions, those not earning free time were given a 20-minute study hall period.

Final Baseline (A)

The procedures in the initial baseline were reinstated.

Data Collection Procedures

The dependent measure used to evaluate the reinforcement procedures was the accuracy of each child's performance in three academic subjects (reading, mathematics, and English). At the end of each 45-minute academic period, the teacher recorded the number of items on each child's worksheet with complete or partial answers. Unanswered items which preceded items with answers were also counted as attempted. Accuracy was then computed by dividing the number of correct answers by the number attempted and multiplying by 100. Children worked at their own rate on open-ended assignments. Each child was given material suitable to his educational level and was allowed to ask for the teacher's assistance if he did not understand the task.

Results

Subject 1 exhibited a mean increase of 20% over his baseline level of reading accuracy during the reading reinforced condition (see Figure 1). This gain in performance was maintained during the chance reinforcement condition. Institution of the final baseline condition did not affect the level of reading performance achieved during the two reinforcement conditions. Although the English answers were not reinforced, English performance showed a mean increase of 33% during the reading reinforced condition (see Figure 2). When the chance reinforcement condition was introduced, there was an initial decrease in English accuracy followed by an increasing trend which stabilized at a level higher than that achieved during the reading reinforced condition. Institution of the final baseline produced a decrease in the level of English performance observed during the two reinforcement conditions. The reading reinforced condition also produced a mean increase in mathematics accuracy of 40% (see Figure 3). During the chance reinforcement condition, mathematics performance appeared to stabilize at a somewhat higher level than was observed during the reading reinforced condition. Mathematics accuracy decreased slightly and showed large variations in the final baseline. The mean values for each subject and condition are presented in Table 2.

The results for Subject 2 also indicated that there was a slight increase in reading

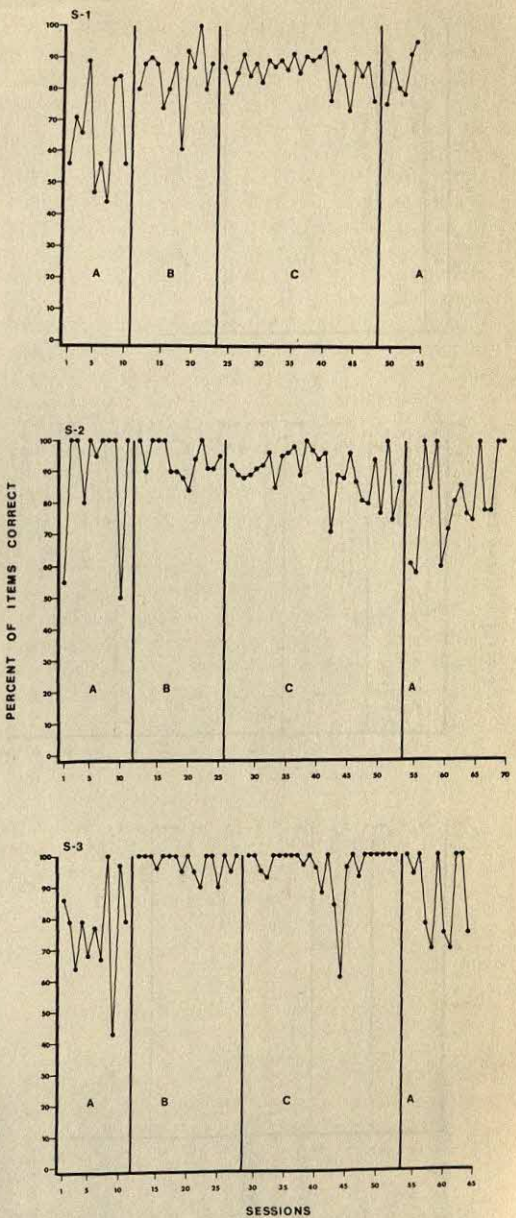


FIGURE 1. Percentage of reading questions answered correctly during each session. (Abbreviations: A = baseline, B = reading reinforced, and C = chance reinforcement.)

accuracy during the reading reinforced condition which was not maintained during the chance reinforcement condition (Figure 1). The downward trend in performance continued during the final baseline. As with

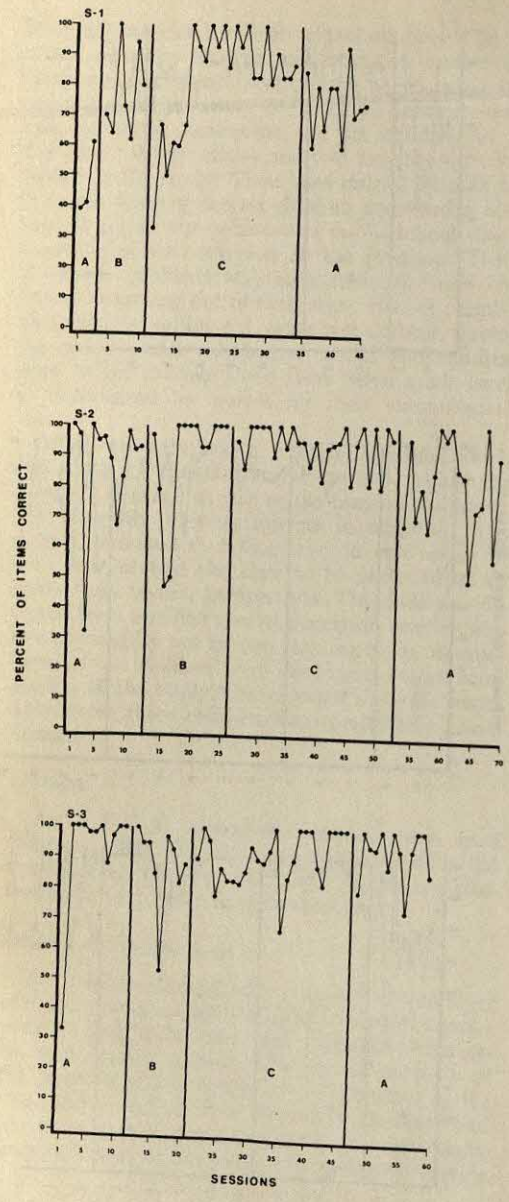


FIGURE 2. Percentage of English questions answered correctly during each session. (Abbreviations: A = baseline, B = reading reinforced, and C = chance reinforcement.)

Subject 1, English performance increased over the initial baseline level during the reading reinforced condition (see Figure 2). Institution of chance reinforcement initially maintained English accuracy at a high level with increased variability during the latter half of the condition. As with reading, these gains were not maintained during the final baseline condition. As noted with English, mathematics performance increased during the reading reinforcement condition to a near perfect level (see Figure 3). Institution of chance reinforcement produced an immediate decrease in accuracy to a lower and more variable rate than that observed during the reading reinforced condition. There was some decrease in performance when the final baseline was instituted.

As with the other two children, Subject 3

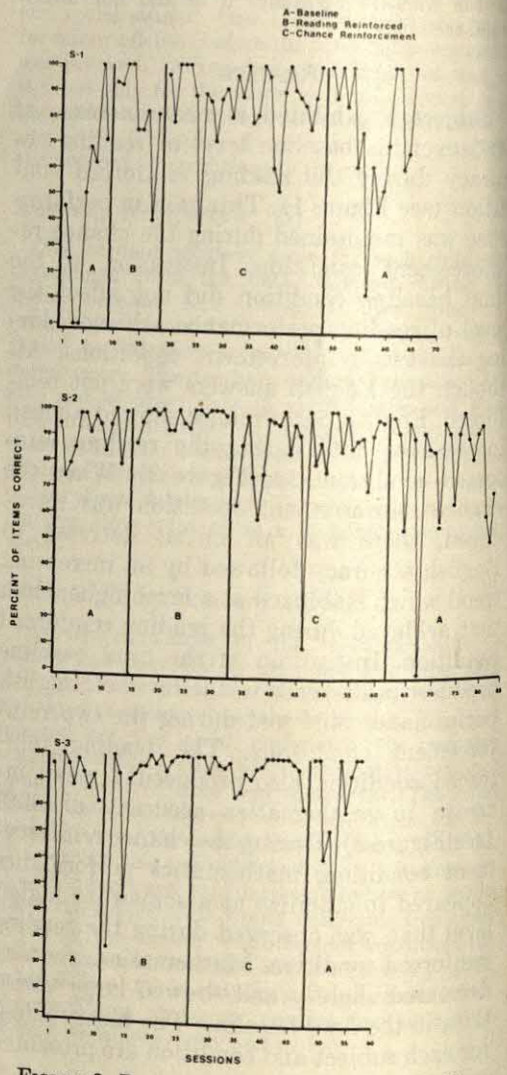


FIGURE 3. Percentage of mathematics problems answered correctly during each session. (Abbreviations: A = baseline, B = reading reinforced, and C = chance reinforcement.)

TABLE 2

MEAN ACCURACY SCORES FOR EACH SUBJECT DURING EACH CONDITION

Academic subject	Baseline 1	Reading reinforced	Chance reinforcement	Baseline 2
Subject 1				
Reading	63.8	84.1	86.6	83.0
English	45.7	79.0	85.2	73.9
Mathematics	49.5	89.0	85.3	74.7
Subject 2				
Reading	89.6	93.0	91.5	81.0
English	84.5	85.5	93.0	81.0
Mathematics	89.8	97.5	85.4	79.0
Subject 3				
Reading	75.8	97.5	97.5	91.9
English	96.0	92.4	88.2	87.6
Mathematics	91.4	92.8	93.4	84.4

exhibited an increase in reading accuracy (22%) during the reading reinforced condition (see Figure 1). This increase was maintained in the chance reinforcement condition. Some disruption in performance was observed during the final baseline. English performance was lower than the baseline level in the reading reinforced condition. Institution of chance reinforcement initially had no effect, but some increase in performance was observed during the latter half of the condition. However, this gain was not maintained during the final baseline condition (see Figure 2). During the reading reinforced condition, mathematics accuracy initially became variable but then stabilized at a somewhat higher level than that achieved during the initial baseline (see Figure 3). The level of performance, largely maintained during the chance reinforcement condition, declined during the final baseline.

The percentage of the sessions during each condition in which the child achieved his criterion level was also analyzed (see Figure 4). Subject 1 achieved his criterion level in reading more frequently during both reinforcement conditions than he did during either baseline. This result was also consistent for the other two academic subjects. Criterion levels of mathematics and English performance were achieved more frequently by Subject 2 during both reinforcement conditions than during baseline. During

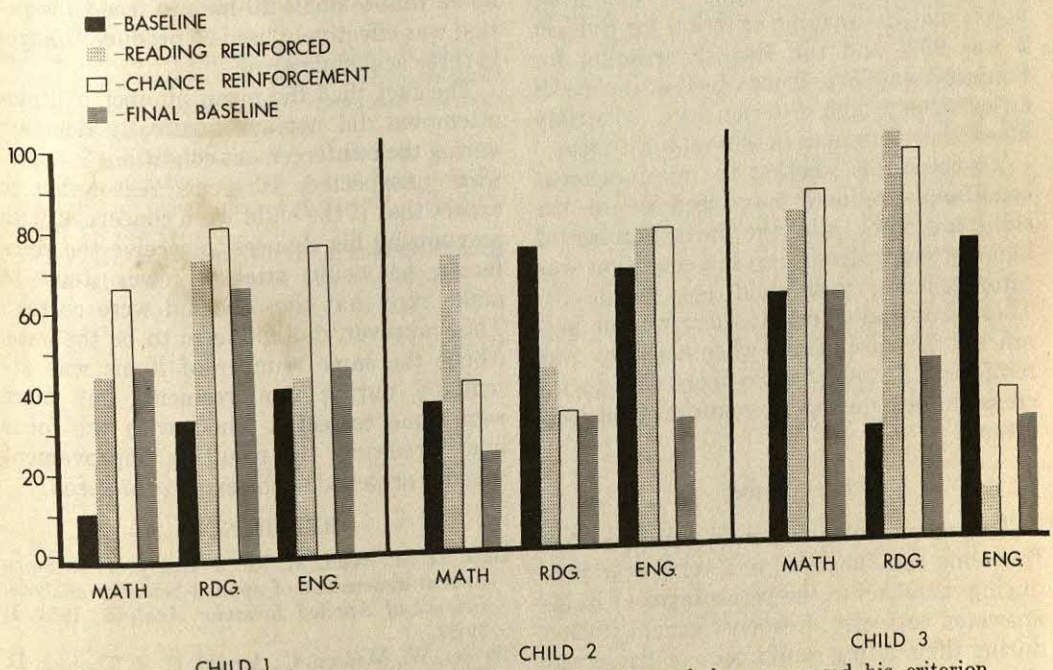


FIGURE 4. Percentage of sessions in which the child equaled or surpassed his criterion level. (Abbreviations: Rdg. = reading, and Eng. = English.)

TABLE 3

MEAN NUMBER OF ITEMS ATTEMPTED DURING THE INITIAL BASELINE AND THE REINFORCEMENT CONDITIONS

Academic subject	Baseline	Reading reinforced	Chance reinforcement
Subject 1			
Mathematics	11.2	31.3	29.3
Reading	26.5	34.8	35.1
English	23.3	28.6	18.6
Subject 2			
Mathematics	98.0	121.5	42.5
Reading	11.4	23.8	31.4
English	35.6	22.8	23.7
Subject 3			
Mathematics	38.3	44.9	44.9
Reading	20.2	17.7	22.3
English	43.4	43.8	37.2

reading, however, the criterion level was achieved less frequently during both reinforcement conditions than it was during the initial baseline. Subject 3 achieved his criterion level in mathematics and reading more frequently during both reinforcement conditions than during baseline. Achievement of his English criterion was less frequent during both reinforcement conditions than during initial baseline. It should be noted that the reading criterion for Subject 2 was 95% and the English criterion for Subject 3 was 98%. It may be that the use of an extremely high criterion level adversely affected performance in these subject areas.

To determine whether the reinforcement conditions produced any decrease in the children's work rate, the mean number of items attempted during each condition was computed for each child (see Table 3). These data indicate that there was no significant decrease in rate when accuracy was reinforced. In many cases there was an increase in rate during the reinforcement conditions.

DISCUSSION

The results of this study indicate that free time can function as a reinforcer producing increases in the percentages of items answered correctly. Response generalization during the reading reinforced condition was observed with both Subject 1 and Subject 2. In the case of Subject 1, generalization was observed in both English and mathematics, while generalization was noted in mathematics but not in English for Subject 2. The lack of any generalized improvement in Subject 3 may have been due to his very high baseline levels in both English and mathematics (96% and 91%, respectively), which in effect left little room for further improvement in these areas. The occurrence of response generalization is of particular interest, since the literature reporting generalization to similar but nonreinforced responses is limited (Lovaas, Koegel, Simmons, & Long, 1973; Schumaker & Sherman, 1970; Wheeler & Sulzer, 1970). Finally, it was apparent that neither the reading reinforced nor the chance reinforcement condition produced consistently differential results.

These results imply that a procedure which reinforces the qualitative aspects of academic performance can be effective in producing increases in academic accuracy. Moreover, these improvements in academic behavior were achieved using free time, a reinforcer which is inexpensive and readily available to the teacher. It should also be noted that a single 20-minute free-time period was effectively used to produce changes in three academic areas.

The fact that the mean number of items attempted did not systematically decrease during the reinforcement conditions is somewhat unexpected. It seems reasonable to expect that if the child were concerned with maximizing his chances to receive the reinforcer, he would attempt fewer items to make sure that those he did were correct. This, however, did not seem to be the case. About the same number of items was attempted during reinforcement, but more were done correctly. Thus, with the focus upon accuracy, the resulting improvement was not achieved at the expense of speed.

REFERENCES

Baer, D. M., Wolf, M. M., & Risley, T. R. Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1968, 1, 91–97.

Becker, W., Madsen, C., Arnold, C., & Thomas, D. The contingent use of teacher attention and praise in reducing classroom behavior problems. *Journal of Special Education*, 1967, 1, 387–407.

Birnbrauer, J., Wolf, M., Kidder, J., & Tague, C. Classroom behavior of retarded pupils with token reinforcement. *Journal of Experimental Child Psychology*, 1965, **2**, 219–235.

Broden, M., Hall, R., Dunlap, A., & Clark, R. Effects of teacher attention and a token reinforcement system in a junior high school special education class. *Exceptional Children*, 1970, **36**, 341–349.

Brown, J., Montgomery, R., & Barclay, J. An example of psychologist management of teacher reinforcement procedures in the elementary classroom. *Psychology in the Schools*, 1969, **6**, 336–340.

Hall, R., Lund, D., & Jackson, D. Effects of teacher attention on study behavior. *Journal of Applied Behavior Analysis*, 1968, **1**, 1–12.

Hewett, F., Taylor, F., & Artuso, A. The Santa Monica Project: Evaluation of an engineered classroom design with emotionally disturbed children. *Exceptional Children*, 1969, **35**, 523–529.

Lovaas, O. I., Koegel, R., Simmons, J. Q., & Long, J. S. Some generalization and follow-up measures on autistic children in behavior therapy. *Journal of Applied Behavior Analysis*, 1973, **6**, 131–165.

Madsen, C., Becker, W., & Thomas, D. Rules, praise, and ignoring: Elements of elementary classroom control. *Journal of Applied Behavior Analysis*, 1968, **1**, 139–150.

Mattos, R., Mattson, R., Walker, H., & Buckley, N. Reinforcement and aversive control in the modification of behavior. *Academic Therapy*, 1969, **5**, 37–52.

Meichenbaum, D., Bowers, K., & Ross, R. Modification of classroom behavior of institutionalized female adolescent offenders. *Behavior Research and Therapy*, 1968, **6**, 343–353.

Nolan, P., Kunzelmann, H., & Haring, N. Behavioral modification in a junior high learning disabilities classroom. *Exceptional Children*, 1967, **33**, 459–467.

O'Leary, K., & Becker, W. Behavior modification of an adjustment class: A token reinforcement program. *Exceptional Children*, 1967, **33**, 637–642.

O'Leary, K., Becker, W., Evans, M., & Saudargas, R. A token reinforcement program in a public school: A replication and systematic analysis. *Journal of Applied Behavior Analysis*, 1969, **2**, 3–14.

Osborne, J. G. Free-time as a reinforcer in the management of classroom behavior. *Journal of Applied Behavior Analysis*, 1969, **2**, 113–118.

Quay, H. C., Werry, J. S. McQueen, M., & Sprague, R. L. Remediation of the conduct problem child in a special class setting. *Exceptional Children*, 1966, **32**, 509–515.

Schumaker, J., & Sherman, J. A. Training generative verb usage by imitation and reinforcement procedures. *Journal of Applied Behavior Analysis*, 1970, **3**, 273–287.

Sidman, M. *Tactics of scientific research*. New York: Basic Books, 1960.

Staats, A., & Butterfield, W. Treatment of nonreading in a culturally deprived juvenile delinquent: An application of reinforcement principles. *Child Development*, 1965, **36**, 925–942.

Thomas, D., Becker, W., & Armstrong, M. Production and elimination of disruptive classroom behavior by systematically varying teacher's behavior. *Journal of Applied Behavior Analysis*, 1968, **1**, 35–46.

Walker, H., Mattson, R., & Buckley, N. Special class placement as a treatment alternative for deviant behavior in children. Eugene: University of Oregon, Department of Special Education, 1969. (Monograph No. 1)

Ward, M., & Baker, B. Reinforcement therapy in the classroom. *Journal of Applied Behavior Analysis*, 1968, **1**, 323–328.

Wheeler, A. J., & Sulzer, B. Operant training and generalization of a verbal response form in a speech-deficient child. *Journal of Applied Behavior Analysis*, 1970, **3**, 139–147.

Wolf, M., Giles, D., & Hall, V. Experiments with token reinforcement in a remedial classroom. *Behavior Research and Therapy*, 1968, **6**, 51–64.

(Received January 29, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 880-888

RETENTION OF PICTORIAL AND VERBAL CONTENT OF A TEXT WITH ILLUSTRATIONS

J. PEECK[1]

University of Utrecht, The Netherlands

Seventy-one fourth graders read an adapted version of a "Rupert Bear" story. In the experimental condition (text with illustrations), pictures and text occasionally provided more or less contradictory information. At retention testing (immediately, after a day, or after a week), the experimental condition produced higher scores than the control (text without illustrations) for questions concerning exclusively pictorial information and for questions concerning correctly illustrated text contents. For questions concerning incongruously presented information, the experimental subjects selected more multiple-choice alternatives representing picture input, while controls preferred alternatives representing textual input. No differences were found for questions covering unillustrated text contents. Neither reading time nor imagery ability were related to retention.

In view of the enormous number of illustrated texts in use (e.g., in education), surprisingly few experimental studies have been performed to investigate the effect of illustrations on learning from reading material. Notable exceptions are some studies by Vernon in the early fifties (e.g., Vernon, 1954), which showed hardly any effect, and a recent series of experiments by Dwyer (for a summary, see Dwyer, 1970), who repeatedly came to the plausible but vague conclusion that some types of visuals are more effective than others in facilitating student achievement of specific educational objectives.

One necessary and obvious approach to this area of study, as yet hardly explored, is the careful distinction between information present (*a*) exclusively in text, (*b*) exclusively in illustration(s), or (*c*) in both, and the investigation of the fate of these three categories of information when retention is tested. Apart from the possibility of furthering insight into the kinds of effects illustrations may have, this line of investigation has the additional advantage of profiting from recent findings in other fields of research where pictures are used as learning material.

From these findings, then, certain effects of illustration seem likely. Many recent studies (e.g., Nickerson, 1965; Shepard, 1967; Standing, 1973) have shown that (complex) pictures are surprisingly well retained. Though most of these studies dealt with recognition of pictures as a whole rather than, for instance, with recall of elements from the pictures, it seems plausible to assume that at least some of this impressive retention may also be found in recall tasks.

When retention of pictures and corresponding names of common objects are compared, pictures again are usually found to do surprisingly well (i.e., they are better retained than the words), this time both on recognition and recall measures and on various kinds of learning tasks (for a review, see Paivio, 1971). In the present context, however, two findings somewhat confuse the issue. First, some studies (Ducharme & Fraisse, 1965; Jenkins, Stack, & Deno, 1969; however, see also Fraisse, 1970) have shown that picture recall may not be as impressive when children are used as subjects. A second difficulty is raised by the

[1] Requests for reprints should be sent to J. Peeck, Psychologisch Laboratorium der Rijksuniversiteit, Varkenmarkt 2, Utrecht, The Netherlands.

outcome that, though pictures are generally found to be better retained than words, the relative retention of pictures *and* words, in comparison to either of these alone, is not altogether clear. Some authors (e.g., Bousfield, Esterson, & Whitmarsh, 1957; Ducharme & Fraisse, 1965) report better retention of items presented in both modes than of items presented verbally, but a recent study of Davies, Milne, and Glennie (1973) indicates that this may not generally be found.

In spite of these obscurities and on the basis of the experiments mentioned so far, one may expect by and large that some retention of pictorial content and some facilitation in retention of pictured text elements may both be obtained when memory of an illustrated instead of an unillustrated text is tested.

Predicting the fate, at retention testing, of text elements not represented in the illustrations is less feasible on the basis of existing research results. On the one hand one could argue that text contents, while not featured in the pictures, might still become associatively connected to the illustrations, thus creating a kind of paired-associate paradigm where the illustration functions as a stimulus term and the text elements as response term(s); the illustration could, in this view, be considered as a kind of "conceptual peg" (Paivio, 1969). Since various authors (for a review, see Paivio, 1969) have shown that in such a set-up, words as response terms are better retained than the same terms in a design where pictures are replaced by corresponding words, one might accordingly expect some facilitation of the retention of the text elements concerned, in comparison to the same elements in a text without any illustrations. Some support for this facilitation can actually be found in a study by Ketcham and Heath (1962), though the claims made by these authors as to the significance of their results are not valid since their analysis of variance was not followed by individual comparisons between means.

A contrary conclusion, however, might also be reached: Illustrations might well be a source of distraction (cf. Samuels, 1970), or they might differentially favor elements from the text that *are* represented in the illustrations, at the cost of the remaining text elements—a phenomenon similar to the effect prequestions may have on the retention of question-irrelevant text content (Peeck, 1970).

Both expectations are not mutually exclusive, however. In an illustrated text, some unillustrated elements from the text may associatively benefit from the illustrations, while the facilitation of these (and of the illustrated elements from the text) may be attained at the cost of other, unillustrated elements.

To study the issues indicated above, subjects in the present experiment were given either an illustrated or an unillustrated text, while the retention test contained questions on elements exclusively present either in the pictures, or in the text, or present in both pictures and text. To gain some insight into the relative contribution of text and pictures on recall, the learning material was so constructed that on some occasions text and picture provided the reader with more or less conflicting information; correspondingly, retention test items (multiple choice) were so designed that a subject could base his answer on the information given by either the text or the picture.

The learning material consisted of an adapted version of a strip cartoon. In order to secure as natural a reading process as possible, subjects (9–10-year-old children) were not given learning instruction nor were they prepared for a retention test. The time they took to read the story was recorded so that a possible effect of the presence of illustrations on reading time might be observed. In order to study the regularity of the reading process, the progress of the children in reading the story was registered at fixed intervals.

Testing for retention took place (*a*) immediately after completion of the story, (*b*) after one day, or (*c*) after a week. In this way it was hoped to gather some data on the course of forgetting of verbal and/or pictorial content, which, as some studies have suggested, may differ (cf. Bahrick & Boucher, 1968; Shepard, 1967).

Method

Material

The learning material was based on the strip cartoon "Bruintje Beer in Dromenland" ("Rupert Bear in the Land of Dreams"). This story seemed a suitable basis for the investigation for several reasons: The pictures are fairly simple yet contain sufficient and sufficiently varied information, and the text constitutes the actual story which can easily be read without the illustrations.

Originally the story consisted of 59 drawings, each of which was accompanied by a text unit of 50–100 words. Of these drawings and text units, 37 were selected for the experiment. The texts were then adapted so as to result in a story with a smaller range in the number of words (70–90) accompanying each drawing. Minor changes in accordance with the objectives of the experiment were then introduced in the text so that pictures plus text contained four categories of information: (*a*) exclusively provided by text (T); (*b*) exclusively provided by pictures (P); (*c*) congruously provided by text and picture, that is, elements from the text correctly pictured in a drawing (P + T); and (*d*) incongruously presented by text and pictures (P ×/ T).

The last category contained, on the one hand, instances where, though text and accompanying picture did not agree, the information provided by the one was not logically incompatible with information given by the other (P/T); for example, on one occasion it is stated in the text that Rupert Bear is walking through a wood, whereas the accompanying picture shows him walking along a country road: The data seem contradictory, but may, with a little good will, be combined; after all, text and picture may deal with different instances of Rupert's behavior. In some other cases, however, the information in the picture was logically virtually irreconcilable with the content of the text (P × T); for example, when, according to the story, a certain character (the hippopotamus) performs a unique action (he pulls the tablecloth off a laid table), whereas the accompanying picture shows the same action being executed by someone else (the king).

The 37 text units, thus (re)constructed, were put under the 37 pictures, xeroxed and made into a new booklet. The unillustrated text units were made into a similar new booklet. Forty multiple-choice questions, each with 4 alternatives, were constructed. There were 8 T questions, 9 P questions, 8 P + T questions, and 15 P ×/ T questions (7 P/T and 8 P × T).

In order to find out the extent to which subjects based their responses on textual or pictorial information, respectively, the 15 P ×/ T questions contained both an alternative corresponding to the text and an alternative corresponding to the picture.

Care was taken to spread the questions from the various categories as evenly as possible over the 37 picture–text units.

All children were given in advance a test of reading ability (a subtest of a Dutch intelligence test, the Interesse-Schoolvorderingen-Intelligentie-Reeks by Snijders, Souren, & Welten, 1968) and a test of imagery ability, the Memory for Designs Test designed by Graham and Kendall (1960; cf. Di Vesta, Ingersoll, & Sunshine, 1971); this made it possible to check the equivalence of the conditions within each school. The Memory for Designs Test, in which the subject has to draw figures shown for five seconds from memory, was chosen because it requires the subject to hold a percept in memory (an activity possibly related to retaining information from illustrations), without the necessity to manipulate (e.g., rotate) it. The items of the test were enlarged and copied on cards for group presentation. Two independent judges scored the reproductions according to the directions given by Graham and Kendall; interrater reliability for all schools together was .85. For each subject, the scores given by the two judges were added together and the resulting scores were used for further computations.

Subjects and Design

Subjects were 71 9–10-year-old children in three fourth grades of three elementary schools (A, B, and C). In each school, about half the children (the experimental condition) were presented with the illustrated text, while the other half (the control condition) read the text without pictures. The children were randomly assigned to the conditions, with the restriction that within each school an attempt was made to get about an equal male-female ratio in each condition. The children in School A were tested for retention immediately after they had read the material; in Schools B and C, unexpected testing took place after one day and one week, respectively.

Neither the reading test nor the Memory for Designs Test showed significant differences between the conditions within a school; on both tests, however, the children in School B obtained generally lower scores than the children in either of the other schools.

On the retention test measures, 3 × 2 (Schools × Conditions) analyses of variance were performed; possible differences between the schools were not further analysed, since, due to the confounding of the time variable with possible differences in intellectual level between the schools, they would be hard to interpret. Differences between experimental and control conditions were tested by *t* tests, one-sided for the scores dealing with retention of purely pictorial material (P in Table 1, and P alternatives in Table 2) and two-sided for all other scores.

Procedure

The experiment was run by two experimenters working simultaneously in order to make the dura-

tion of the experiment in a school as short as possible and to prevent the children from discussing the learning material. The children were tested in groups of 4 or 5, homogeneous as to the condition; they were seated at separate tables and told they would be given a booklet to read the way they would do it when reading for pleasure at home. It was stressed in the instruction that there was no need to hurry, everybody could read at his own rate. The children were then given a booklet and an empty sheet (to draw on if they had finished before the others), and they were told they could start reading. The experimenter noted at two-minute intervals which page of the booklet each child had reached.

After all the children had finished the story and the booklets had been collected, the children were asked, by means of a questionnaire, if they had read "Rupert Bear" books before, and if they knew this particular one. (On the basis of the responses to these questions, six children who presumably had read the story before were removed from the original sample.)

Sheets for the 40-item retention test were then handed out to the subjects in School A; in Schools B and C, testing took place after one and seven days, respectively. The test was introduced as a test of "what you remember of what you have read"; the multiple-choice principle was explained to the children, and they were asked to answer all questions and to guess if they did not know the answer.

Results

Table 1 deals with the retention of unillustrated (T) and correctly illustrated (P + T) text contents and of information exclusively shown in the pictures (P).

In all schools, retention of T contents appears to be somewhat higher in the experimental than in the control condition, though the overall effect is not significant ($F = 2.30$, $df = 1/65$, $.10 < p < .20$). The presence of the illustrations thus seems to have had a slight facilitative effect on the retention of unillustrated elements from the text. The differences between the schools are significant ($F = 3.25$, $df = 2/65$, $p < .05$), but no further analysis or interpretation of this finding will be attempted, for reasons stated above; there was no Schools × Conditions interaction ($F < 1.0$).

Subjects in the experimental condition retained significantly more P + T contents than subjects in the control condition ($F = 11.37$, $df = 1/65$, $p < .005$). As Table 1 shows, further analyses within the schools reveal significant differences in Schools B and C but no significant difference in School A. This suggests that the facilitative effect of illustration on the retention of text elements is primarily a phenomenon of delayed retention testing. Here too the differences between the schools were significant ($F = 5.49$, $df = 2/65$, $p < .01$), whereas the interaction was not ($F = 1.38$, $df = 2/65$, ns).

Finally, the P questions were, as could be

TABLE 1

Mean Retention and Standard Deviations of Unillustrated Text Contents (T), Picture Contents (P), and Correctly Illustrated Text Contents (P + T) at Immediate Testing, Testing after One Day, and Testing after One Week

Contents	School A (immediate)			School B (one day)			School C (one week)		
	Exp	Con	t difference	Exp	Con	t difference	Exp	Con	t difference
n	10	11		16	14		9	11	
T									
$\bar{X}$	6.00	5.45	1.00	4.81	4.21	.95	5.22	4.55	.87
SD	1.18	1.30		1.59	1.86		1.81	1.67	
P									
$\bar{X}$	5.60	4.36	2.85**	6.25	3.29	6.06**	5.56	3.00	2.84**
SD	1.02	.98		1.56	1.03		2.06	1.95	
P + T									
$\bar{X}$	4.70	4.27	.69	5.44	3.43	2.93**	3.89	2.09	2.49*
SD	1.42	1.42		1.84	1.92		1.52	1.68	

Note. Abbreviations: Exp = text with illustrations; Con = text without illustrations.

* $p < .05$.

** $p < .01$.

TABLE 2

MEAN SELECTION OF ALTERNATIVES CORRESPONDING TO TEXT (T) OR TO PICTURES (P), ON MULTIPLE-CHOICE QUESTIONS DEALING WITH INFORMATION PRESENTED MORE (P × T) OR LESS (P/T) INCONGRUOUSLY BY TEXT AND PICTURES

Question alternative		School A (immediate)			School B (one day)			School C (one week)		
		Exp	Con	t difference	Exp	Con	t difference	Exp	Con	t difference
P/T										
Picture	$\bar{X}$	2.90	1.09	3.75*	2.56	1.14	4.22*	2.56	1.27	2.25*
	SD	1.30	.90		.86	.99		1.26	1.29	
Text	$\bar{X}$	3.50	4.09	.89	3.31	4.14	1.78	3.33	3.73	.76
	SD	1.50	1.50		.98	1.55		.94	1.29	
P × T										
Picture	$\bar{X}$	4.10	1.27	4.88**	3.81	2.79	1.90*	4.78	2.63	3.63**
	SD	1.51	1.14		1.38	1.21		1.93	.98	
Text	$\bar{X}$	3.00	4.64	3.34**	2.75	2.79	.70	2.11	2.64	.93
	SD	1.18	1.07		.83	1.15		1.52	.98	
Total P ×/T										
Picture	$\bar{X}$	7.00	2.36	5.64**	6.38	3.93	3.03**	7.33	3.64	4.33**
	SD	2.10	1.67		2.03	1.80		2.21	1.61	
Text	$\bar{X}$	6.50	8.73	2.41*	6.06	6.93	1.20	5.44	6.36	1.26
	SD	2.06	2.09		1.60	2.31		1.78	1.49	

Note. Abbreviations: Exp = text with illustrations; Con = text without illustrations.
* $p < .05$.
** $p < .01$.

expected, answered correctly significantly more often in the experimental than in the control condition ($F = 37.12$, $df = 1/65$, $p < .001$), a superiority found, as the t tests show, in all schools. Differences between the schools were not significant ($F = 1.29$, $df = 2/65$, ns) nor was the interaction ($F = 1.97$, $df = 2/65$, ns). Interestingly, the retention of pictorial content shows, as can be seen in Table 2, little deterioration as the interval between learning and testing increases. However, as was said before, due to the lack of equivalence of the schools, this finding should be viewed with caution.

Table 2 shows the result on items dealing with information more (P × T) or less (P/T) incongruously presented in text and picture.

On both the P/T and P × T questions, subjects in the experimental condition chose significantly more often than control subjects the alternative corresponding to the picture content (P alternative) ($F = 30.68$ and $F = 37.64$, respectively, with $df = 1/65$ and $p < .001$ in each case). There are no significant differences between the schools ($F < 1.0$ and $F = 2.36$, respectively, with $df = 2/65$), nor were the Schools × Conditions interactions significant ($F < 1.0$ and $F = 2.54$, with $df = 2/65$). As can be seen in Table 2, in all schools, t tests between experimental and control conditions were significant on at least the .05 level.

On the alternatives corresponding to the information provided by the text (T alternatives), a significant difference between the conditions was only found for the P × T questions ($F = 6.98$, $df = 1/65$, $p < .05$); for these T alternatives, there were highly significant differences between the schools ($F = 9.78$, $df = 2/65$, $p < .001$), while the interaction also reached significance ($F = 3.36$, $df = 2/65$, $p < .05$). As is shown in Table 2, only on immediate testing (School A), control subjects chose significantly more T alternatives on P × T questions. On the other category of questions (P/T), the difference in choice of T alternatives between experimental and control conditions was almost significant ($F = 3.38$, $df = 1/65$, $.05 < p < .10$); here neither the differences between the schools nor the interaction were significant (both times $F < 1.0$).

So far the results on the P/T and P × T questions have been treated separately, since, as was explained above, these cate-

gories differed in the degree of conflict between textual and pictorial information involved. However, as it is not known whether subjects actually experienced this distinction, it is perhaps advisable to also analyse these categories together (P×/T, see Table 2). Analysis of variance of the P alternatives then reveals a highly significant difference between the conditions ($F = 56.81$, $df = 1/65$, $p < .001$); neither differences between the schools nor interaction reach significance ($F = 1.10$ and $F = 1.77$, respectively, with $df = 2/65$). The t test of the difference between the experimental and control conditions is significant beyond the .01 level for each school (see Table 2). On the T alternatives, a significant difference between the conditions is found ($F = 7.82$, $df = 1/65$, $p < .01$); this time the schools differ significantly as well ($F = 4.40$, $df = 2/65$, $p < .05$), while there is no significant interaction ($F < 1.0$). Additional t tests show that on immediate testing only, a significant difference between the conditions is obtained (in favor of the control condition).

Taken together these results indicate that the exposure of subjects in the experimental condition to both pictures and text had a powerful effect on their behavior on the retention test: They chose the alternative corresponding to the pictorial information far more frequently than the subjects who had only seen the text, for whom, it should be realized, the P alternatives were merely one of three (incorrect) response alternatives. This preference is probably responsible for the results on the T alternatives, which the experimental subjects selected less frequently than the control subjects.

It is worth noticing further that with increasing retention intervals, the frequency of selection of P alternatives remains about the same, whereas the frequency of selection of T alternatives shows a steady decline.

The children showed considerable variation in reading time. In two of the schools, subjects in the experimental condition completed the story on the average somewhat faster than the controls and in the third school somewhat slower. However, none of these differences was significant (in Schools A, B, and C, $t = -1.85$, 1.66, and -1.39, respectively, all $ps > .05$). There is, therefore, no indication that the presence of the pictures affected the inspection rate of the children in any systematic way.

In order to explore the regularity of the reading rate over the story, in both conditions in each school, product-moment correlations were computed between total reading time and the number of pages read after 8 minutes. This point was chosen because the average reading time of all children was about 16 minutes, and it therefore could be assumed that after 8 minutes the subjects were about halfway through the story. For the experimental and control conditions, respectively this correlation was as follows: School A, $-.86$ and $-.97$; School B, $-.82$ and $-.83$; and School C, $-.92$ and $-.88$. Inspection rate thus appears to have been very regular and, in this respect too, it seems unaffected by the presence or absence of pictures.

Correlations were also computed between the various retention measures on the one hand, and reading time, the scores on the reading test, and the scores on the Memory for Designs Test on the other. No systematic correlations were found for either reading time or imagery scores. Only the scores on the test of reading ability showed consistently positive and somewhat substantial correlations with scores on T questions, and T alternatives on P×/T questions.

Discussion

The results of this experiment clearly show that the presence of illustrations had several effects on retention: On questions covering exclusively pictorial content (P) and on questions dealing with information uniformly presented by both text and illustrations (P + T), subjects who read the illustrated text scored significantly higher than subjects who read the text without illustrations; on items covering information incongruously presented by text and illustration (P/T and P × T), they checked the alternative corresponding to the pictorial content far more frequently than the subjects who only read the text, whereas the latter tended to select the alternative corre-

sponding to the text more often. Though no significant difference was obtained for items testing retention of information presented through the text only (T), an indication of some facilitation in the illustrated text condition was found in all schools.

In relating these results to educational practice, it should be noted that the learning material used was probably more conducive to pictorial retention than illustrated textbooks used in education generally are. Though the text of this strip cartoon carried the actual story, the pictures continuously provided the reader with some visualization of its contents, thus reinforcing the tendency, already natural in reading strip cartoons, to pay attention to the illustrations. As a result, the retention of the pictorial material in the present experiment may be somewhat inflated. However, there seems as yet little theoretically founded reason to expect radically different phenomena when retention of illustrated textbooks (for this age group) is investigated with methods similar to the ones applied in this study, although the effect of illustration will probably turn out to be less marked.

Due to differences between the schools, no definitive conclusions can be drawn concerning the existence of systematic differences in the course of forgetting of pictorial, as compared with verbal content. There are, however, several indications in the data that retention of pictorial content might suffer less from an increase of the interval between learning and testing than retention of verbal content. This would be in agreement with results obtained in the studies, referred to above, which showed the impressive, and relatively stable, long-term retention of pictures.

The findings on the items dealing with incongruously presented information require some further comment. They show that subjects in the illustrated text condition tended, on the average, to select more alternatives congruent with the pictures than alternatives congruent with the text. This raises the question whether subjects, when selecting P alternatives, were consciously basing their answers on pictorial information, and whether they were (either on testing or before, while reading) aware of contradictions between text and pictures. In this respect, it should be noted that the children in the illustrated text condition were relatively free to base their answer on either source of information, relatively free because the retention test was introduced as a test "to see what you remember of what you have read." Furthermore, in educational practice, pictures in textbooks often tend to be neglected as a legitimate source of knowledge, and testing often deals exclusively with verbally presented information. It may therefore be argued that, also because of the scholastic context, the children in the present experiment considered the retention test as primarily a test of their retention of the text. Accordingly, the children in the experimental condition probably tried to base their answer either on what they thought they had read in the text, or, in the case of conscious reference to the pictures, assumed that text and pictures had provided the same information. This would imply that little awareness of contradiction was raised in initial learning, or, if raised, was on testing largely forgotten or overridden, due to the superior retention of pictorial content. This interpretation is supported by the observation that the children did not show any sign of awareness of the presence of conflicting information in the reading material, either on initial reading or on testing.

In order to obtain more data on these matters, several lines of investigation may be tried. One line, for example, would be to emphatically instruct the children to base their answers on the retention test exclusively on information provided by the text. Another line would be to see whether subjects are capable of distinguishing between pictures and text as the source for answering test items; this might give some indication of the way the information from both sources is stored (cf. Sasson, 1971). Apart from this, it would be useful to find out, for example, by postexperimental questioning, to what extent children perceived inconsistencies in the learning material.

It is not clear how the absence of systematic differences in reading time between the

illustrated and unillustrated text conditions should be interpreted. It could mean that contrary to what was suggested above, the children who read the illustrated text paid relatively little attention to the pictures, otherwise their total inspection time would, on the average, probably have exceeded the reading time in the other condition (cf. Dwyer, 1968). It is more likely though, that the time spent on the illustrations was compensated for by a more rapid reading of the text itself. This could be due to motivational causes (reading was more fun, more curiosity was raised, etc.), but it is also possible that the story could be more easily understood, and therefore read more rapidly, when accompanied by pictures. The latter assumption is in accordance with the small but consistent difference in retention of unillustrated text contents (T), shown in Table 1.

The absence of correlations between the imagery test scores and the retention test scores is not too surprising. It may indicate that bringing to mind pictures and text of a story read up to a week before, involves quite different aspects of imagery ability than having to copy from memory geometrical figures, seen a few seconds before.

Finally, mention should be made of an additional analysis of the data not originally intended. As some recent studies (e.g., Marks, 1973) suggest, there may be systematic sex differences involved in picture-memory experiments. It seemed therefore worthwhile to find out whether differences of this kind might have occurred in the present experiment. To this purpose, the results of the boys and girls were compared (all schools combined) for the experimental and for the control condition. No differences were found except, in the experimental condition, on items dealing with incongruously presented information (P×/T). On these items, girls tended to choose answers corresponding to the pictorial content (P alternatives) more often than boys ($t = 1.70$, $df = 33$, $.05 < p < .10$), whereas boys were more inclined than girls to select answers corresponding to textual information ($t = 2.00$, $df = 33$, $.05 < p < .10$). This might be taken as support for the suggestion by Ernest and Paivio (1971) that "in some tasks females 'use' imaginal processes to facilitate recall whereas males do not [p. 71]."

REFERENCES

Bahrick, H. P., & Boucher, B. Retention of visual and verbal codes of the same stimuli. *Journal of Experimental Psychology*, 1968, **78**, 417–422.

Bousfield, W. A., Esterson, J., & Whitmarsh, G. A. The effects of concomitant colored and uncolored pictorial representations on the learning of stimulus words. *Journal of Applied Psychology*, 1957, **41**, 165–168.

Davies, G. M., Milne, J. E., & Glennie, B. J. On the significance of "double encoding" for the superior recall of pictures to names. *Quarterly Journal of Experimental Psychology*, 1973, **25**, 413–423.

Di Vesta, F. J., Ingersoll, G., & Sunshine, P. H. A factor analysis of imagery tests. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 471–479.

Ducharme, R., & Fraisse, P. Etude génétique de la mémorisation de mots et d'images. *Canadian Journal of Psychology*, 1965, **19**, 253–261.

Dwyer, F. M. The effectiveness of visual illustrations used to complement programed instruction. *Journal of Psychology*, 1968, **70**, 157–162.

Dwyer, F. M. Exploratory studies in the effectiveness of visual illustrations. *Audio Visual Communication Review*, 1970, **18**, 235–249.

Ernest, C. H., & Paivio, A. Imagery and sex differences in incidental recall. *British Journal of Psychology*, 1971, **62**, 67–72.

Fraisse, P. La verbalisation d'un dessin facilite-t-elle son évocation par l'enfant. *Année Psychologique*, 1970, **70**, 109–122.

Graham, F. K., & Kendall, B. S. Memory-for-Designs Test: Revised general manual. *Perceptual and Motor Skills*, 1960, **11**, 147–188.

Jenkins, J. R., Stack, W. B., & Deno, S. L. Children's recognition and recall of picture and word stimuli. *Audio Visual Communication Review*, 1969, **17**, 265–271.

Ketcham, C. H., & Heath, R. W. Teaching effectiveness of sound with pictures that do not embody the material being taught. *Audio Visual Communication Review*, 1962, **10**, 89–93.

Marks, D. F. Visual imagery differences in the recall of pictures. *British Journal of Psychology*, 1973, **64**, 17–24.

Nickerson, R. S. Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology*, 1965, **19**, 155–160.

Paivio, A. Mental imagery in associative learning and memory. *Psychological Review*, 1969, **76**, 241–263.

Paivio, A. *Imagery and verbal processes*. New York: Holt, Rinehart & Winston, 1971.

Peeck, J. Effect of prequestions on delayed retention of prose material. *Journal of Educational Psychology*, 1970, **61**, 241–246.

Samuels, S. J. Effects of pictures on learning to read, comprehension and attitudes. *Review of Educational Research*, 1970, **40**, 397–407.

Sasson, R. Interfering images at sentence retrieval. *Journal of Experimental Psychology*. 1971, **89**, 56–62.

Shepard, R. N. Recognition memory for words, sentences and pictures. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 156–163.

Snijders, J. T., Souren, C. J. M. H., & Welten, V. J. *Interesse - Schoolvorderingen - Intelligentie-Reeks*. Groningen: Wolters, 1968.

Standing, L. Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, 1973, **25**, 207–222.

Vernon, M. D. The instruction of children by pictorial illustration. *British Journal of Educational Psychology*, 1954, **24**, 171–179.

(Received January 29, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 889-894

SERIAL MEMORY SPAN THRESHOLDS OF NORMAL AND MENTALLY RETARDED CHILDREN[1]

ALFRED A. BAUMEISTER[2]

George Peabody College for Teachers

Memory span thresholds were obtained for second-grade, sixth-grade, and retarded children over seven classes of materials in both auditory and visual modalities. One of the tasks involved a procedure designed to force "chunking." The major conclusions were as follows: (*a*) a clear developmental effect was observed which generalized across materials and modalities; (*b*) the retarded subjects were consistently inferior to normal children; (*c*) auditory presentation produced better recall than visual presentation; (*d*) relatively high intercorrelations were observed among the various thresholds for all subject groups, but particularly for the older normal children; (*e*) forced chunking produced longer spans, but not differentially across groups; and (*f*) retarded subjects, as contrasted with normal children, may employ different processing strategies.

A task most commonly used to measure short-term retention capacity is the venerable memory span test, particulary memory for digits. This task is assumed to be a good measure of the memory component of intelligence (e.g., Scott & Scott, 1968) and, in fact, digit spans correlate well (.60 to .75) with Stanford-Binet and Wechsler IQs. Additionally, sequential short-term memory processes have figured prominently into various theoretical accounts of retarded behavior (e.g., Ellis, 1970; Jensen, 1970; Spitz, 1973).

The process which mediates the ability to remember a subspan sequence of items has been described by a variety of terms including *apprehension span*, *short-term memory*, and *channel capacity*, among others. By whatever name one wishes to call it, there are some descriptive facts about this process which are especially pertinent to theoretical issues in the area of mental retardation: (*a*) digit span thresholds are positively correlated with mental age; (*b*) immediate memory span depends upon the type of material (Cavanagh, 1972); (*c*) retarded individuals typically exhibit a smaller memory span (digit) capacity than normal individuals (Spitz, 1973); (*d*) modality effects are commonly observed in short-term memory tasks (Madigan, 1971); and (*e*) temporal or spatial grouping of items in a sequence influences memory span (Spitz, 1973).

The major objective of this study was to investigate concurrently these various effects among normal and retarded children. Of particular concern were interactions involving developmental variables, type of material, and modality. Also of interest were correlations of memory span for different types of materials.

Method

Subjects

All subjects were obtained from a public elementary school serving a predominantly white, middle-class population. Three groups of 15 children were randomly selected from a second grade, a sixth grade, and a special class for educable mental retardates. The mean age of the second graders was 7.9 years with a standard deviation of .5 years. The corresponding descriptive statistics for the sixth-grade children were 11.9 and .7 years, respectively.

The mean Stanford-Binet IQ for the retarded children was 71.7 (SD = 5.1). Their chronological age was 11.0 years with a standard deviation of 1.3 years. Mean mental age, obtained for each retarded subject, was 7.8 years (SD = 1.3 years). So, roughly speaking, the retarded children were matched with the second-grade children in terms

[1] This research was supported by PHS Grants HD02588 and HD00973.

[2] Requests for reprints should be sent to Alfred A. Baumeister, Department of Psychology, Box 6234, University of Alabama, University, Alabama 35486.

of mean mental age and with the sixth-grade children in terms of mean chronological age.

Sex distribution was approximately equal among the three groups. There were 9, 7, and 7 girls, respectively, among the second-grade, sixth-grade, and special-class children.

Design and Procedures

Memory span thresholds were obtained for seven classes of materials (described below) in both visual and auditory modalities. This procedure resulted in 14 measures for each subject.

The method for determining thresholds was similar to the procedures used with respect to the Digit Span Test contained in the Wechsler Intelligence Scale for Children (WISC). A sequence of given length was presented twice. The number of items in a sequence was increased one at a time until the child failed two sequences of a given length. Memory span was defined as the longest sequence the subject could recall, without error, on two successive occasions. Two or three practice trials were given prior to each testing session.

The examiner started each testing session with a two-item sequence. Items were presented at the rate of one per second. A successive presentation procedure was used to render auditory and visual presentations comparable. The auditory materials were spoken by the examiner. A slide projector was used for the visual presentation.

The order of testing, with respect to both material and modality, was random for each subject. After completing the initial series of 14 tests, the entire procedure was repeated for each subject. This provided test–retest reliabilities of the threshold for each type of material under auditory and visual presentation.

Materials

Seven classes of material were used to obtain immediate memory span thresholds: digits, letters, three-letter words, animal pictures, colors, mixed (letters and digits), and paired associates of letters and digits. Materials were presented both visually and orally.

1. The digit span lists were constructed to avoid common dates (e.g., 1973) and sequences which follow each other in the usual counting order.

2. Letter sequences were arranged randomly with the exception that care was taken to avoid the spelling of words or combinations of letters that follow each other in the alphabet.

3. The animal lists were composed of the following: tiger, monkey, sheep, horse, turkey, rabbit, cow, chick, and pig. Preliminary to testing, each subject was asked to identify the animal pictures and names.

4. The word lists were constructed from three-letter words chosen from a second-grade reading book. These included bee, car, pot, egg, hat, man, boy, cat, nut, pig, cow, pin, rat, dog, toy, sun, top, and day. Subjects were asked to identify each word before testing.

5. The colors were as follows: red, orange, black, brown, blue, yellow, green, and purple. The visual presentation was accomplished by placing colored tapes on clear slides. Again, subjects were asked to identify the colors before they were tested.

6. The mixed class of material consisted of letters and digits. Common associations were avoided. When the sequence contained an even number of terms, digits and letters were equally represented in random order. When the sequence included an odd number of items, the lists were counterbalanced such that letters and numbers were presented equally often across lists.

7. In an effort to include a task that forced some organization upon the material, the association memory span test was devised. The associative tasks consisted of letters (stimuli) and digits (responses) shown or read together, for example, Z–5, P–3, M–2, and V–9. Common associations were avoided. Visual presentation consisted of simultaneous presentation of both terms of a pair, while in the auditory mode a successive presentation was made.

Results

A variety of procedures have been employed to score memory spans, ranging from correct or incorrect for the entire sequence to a weighted system that gives credit for partially accurate responses. These various scoring methods produce highly correlated results for children and retardates (Hawkins & Baumeister, 1965). In view of this fact and because the focus of this study was upon memory *thresholds*, the dependent measure employed here was the longest sequence the subject was able to recall correctly twice in succession.

Table 1 presents mean thresholds for each of the three groups by type of material and modality. These scores are averaged over the two testing sessions.

Test–retest reliability coefficients are presented in Table 2. Although no statistical comparisons were made of the reliability coefficients, it is clear that they are generally somewhat higher for the sixth-grade children than for either of the other two subject groups.

Preliminary analysis were conducted to determine whether the sex variable was related to recall thresholds. No significant difference was obtained for any comparison and, for all further analyses of these data, the sex variable was ignored.

A mixed effects $3 \times 7 \times 2 \times 2$ (Group $\times$

TABLE 1
MEAN THRESHOLDS

Material	Group[a]		
	Mentally retarded	2nd grade	6th grade
Digit			
Visual	3.6 (.9)	4.3 (1.0)	5.9 (1.3)
Auditory	3.6 (.8)	4.8 (.8)	5.9 (1.1)
Color			
Visual	3.1 (.8)	3.9 (.7)	5.0 (1.1)
Auditory	3.5 (.8)	4.7 (.9)	5.2 (.9)
Letter			
Visual	2.8 (.8)	3.8 (.9)	5.0 (1.1)
Auditory	3.4 (.9)	4.5 (.9)	5.4 (1.1)
Association			
Visual	1.7 (.4)	2.2 (.6)	2.8 (.5)
Auditory	2.1 (.5)	2.6 (.5)	3.0 (.7)
Word			
Visual	2.6 (.5)	3.6 (1.1)	5.0 (1.3)
Auditory	3.6 (.8)	4.1 (1.0)	4.8 (.8)
Mixed			
Visual	3.3 (.8)	4.0 (1.0)	5.4 (1.2)
Auditory	3.6 (.7)	4.5 (1.0)	5.5 (1.1)
Picture			
Visual	3.0 (.7)	4.0 (.8)	4.8 (1.1)
Auditory	3.4 (.7)	4.1 (.5)	5.3 (1.0)

[a] Standard deviations are given in parentheses.

Material × Modality × Test–Retest) analysis of variance was applied to the threshold scores. Significant main effects were found for all variables (groups: $F = 23.5$, $df = 2/42$, $p < .001$; type of material: $F = 157.7$, $df = 6/252$, $p < .001$; modality: $F = 35.0$, $df = 1/42$, $p < .001$; test–retest: $F = 49.0$, $df = 1/42$, $p < .001$). Under all conditions, the sixth graders were superior to the second graders, who were, in turn, superior to the special-class children. With few exceptions, auditory presentation produced better recall than visual presentation. Performance was generally better on the second occasion of testing than on the first. Digit span thresholds were highest and, of course, the association spans were lowest. The mixed class of materials, including digits and letters, ranked second for all groups. Overall, none of the differences between thresholds for colors, letters, words, or pictures was marked.

However, some significant interactions were observed, including Group × Type of Material ($F = 5.0$, $df = 12/252$, $p < .001$), Group × Test–Retest ($F = 3.9$, $df = 2/42$, $p < .03$), Modality × Test–Retest ($F = 10.8$, $df = 1/42$, $p < .01$), Type of Material × Test–Retest ($F = 2.5$, $df = 6/252$, $p < .03$), Group × Type of Material × Modality ($F = 2.5$, $df = 12/252$, $p < .004$).

These interactions were further analysed by the Tukey Q procedure. Following are some of the clearer findings that emerged from these subanalyses: (*a*) auditory thresholds were generally higher than visual thresholds, more so on the first occasion of testing than on the second; (*b*) the test–retest effect was greater for the visual modality than for the auditory modality but significant in both cases; (*c*) generalizing across materials and modalities, the practice effect was greater for the sixth graders, least for the retarded children, and significant for all groups; and (*d*) ignoring the association spans, the retarded children exhibited virtually identical thresholds for all material presented in the auditory modality. Given visual presentation, however, the retarded subjects displayed fairly large threshold differences across the various types of materials. The sixth graders, on the other hand,

TABLE 2
TEST–RETEST RELIABILITY COEFFICIENTS FOR EACH GROUP BY TYPE OF MATERIAL AND MODALITY OF PRESENTATION

Span	Group		
	Mentally retarded	2nd grade	6th grade
Digit			
Visual	.67	.44	.82
Auditory	.53	.51	.78
Color			
Visual	.35	.18	.75
Auditory	.82	.46	.33
Letter			
Visual	.83	.44	.63
Auditory	.79	.68	.80
Association			
Visual	.00	.60	.65
Auditory	.59	.87	.75
Word			
Visual	.48	.76	.93
Auditory	.53	.68	.61
Mixed			
Visual	.57	.38	.74
Auditory	.75	.61	.71
Picture			
Visual	.64	−.04	.77
Auditory	.30	.12	.63

revealed threshold differences for the various types of material under both visual and auditory modalities. The second-grade subjects were intermediate in these respects.

Correlations were computed among thresholds for the various classes of materials within each of the subject groups. Considering the sample sizes in each group relative to the number of variables measured for each subject, it was not feasible to perform a factor analysis upon these intercorrelations. However, it is possible to look comparatively at some of the general patterns that emerged from the correlations. The average correlations of all threshold spans were .50, .63, and .76 for the retarded subjects, and the second graders, and the sixth graders, respectively.

Another way of looking at these results is to consider the correlations of within-spans modality as compared with between-spans modalities. These correlations are presented in Table 3. In general, it appears that the relationships among memory span thresholds for various types of material are somewhat higher when the materials are presented within the same modality.

One of the tasks, in particular, explicitly incorporates reading skills—words presented visually. It was on this task that the retardates and second graders displayed their relatively weakest performance (excluding the association thresholds). The sixth graders, on the other hand, performed as well with words as they did with colors, letters, and pictures. The correlation between visual word spans and span thresholds for other types of visual material were .75, .65, and .24 for the sixth graders, second graders, and the special-class pupils, respectively. This was a particularly reliable task for the sixth graders (r = .93).

TABLE 3

INTERCORRELATIONS FOR VISUAL, AUDITORY, AND VISUAL–AUDITORY TASKS

Group	Visual	Visual-auditory	Auditory	$\bar{X}$
Mentally retarded	.42	.46	.61	.50
2nd grade	.72	.57	.60	.63
6th grade	.78	.72	.79	.76
$\bar{X}$	.64	.58	.67	

DISCUSSION

These results clearly show a developmental trend in memory span capacity for a large variety of materials. In every instance of the seven classes of materials, the sixth-grade children exhibited significantly higher thresholds than either of the other groups. This is not an entirely unexpected finding, in that growth functions for memory span have been observed previously for both normal and retarded subjects (e.g., Korst & Irwin, 1968). The obvious inferiority of the retarded subjects is of considerable interest, particularly when the comparison is made against the normal subjects of matched mental age. Again, for every type of material, the special-class students had lower thresholds. The finding that this deficiency applied for all types of materials in both the auditory and visual modality suggests that the relationship between memory span and IQ is a fairly general phenomenon.

The source of these developmental trends, however, is not readily determined from the present study which was designed primarily for parametric purposes. One important issue is whether the developmental differences observed here represent differential encoding or retrieval capacities. There is some evidence to suggest that encoding factors may be the source of the differential performance among the three subject groups. The older and/or brighter subjects may have employed more sophisticated grouping or chunking strategies (MacMillan, 1970; Spitz, 1973).

The associative spans were considered to be a special case of the mixed-class materials. One way of looking at the associative task is that it requires the subject to "chunk" or organize the individual items into groups of two. So, by doubling the number of associations that defined each subjects' threshold, a measure of total items remembered in series was obtained. A comparison of the mixed-class spans and the associative spans enables us to determine the effect that forced chunking has upon memory span.

Because of the low reliability of the visual association task for the special-education children, only the auditory tasks were compared. The advantage of the associative over the mixed spans was .7, .6, and .5 items for the special-class, second-grade, and sixth-grade children, respectively. While the overall difference between types of spans was significant, none of the differences between subject groups was significant, indicating that all groups benefited about equally from this type of associative chunking.

Another general finding was that auditory presentation produced higher thresholds than visual presentation of the material. Again, the study was not designed to elucidate the processes that might be implicated with respect to modality effects; but, in consideration of the oral recall requirement, we may presume that less transformation of the material was required under the auditory presentation.

Modality effects favoring auditory input have been observed previously in short-term memory experiments. A number of hypotheses have been advanced to account for this phenomenon. For instance, it has been suggested that auditory stimuli are encoded more rapidly into short-term memory (Laughery & Pinkus, 1966). Others have argued for separate iconic and echoic memory stores with input remaining active longer in the latter (e.g., Craik, 1969). In any case, this is a difference that seems to occur across various types of inputs, with the possible exception of digits. The magnitude of the modality effect seems to be related to developmental status. The significant interaction involving Groups × Modality × Materials was, in part, attributable to a much more comparable performance of the sixth-grade children for auditory and visual presentation. Differences between the second graders and the retardates were most pronounced in the auditory modality for relatively simple stimuli (digits, colors, letters, mixed).

It should be noted that while span thresholds were obtained for seven different classes of materials, the distinction seems to be more obvious and perhaps more meaningful in relation to the visual modality. Under the auditory condition, names of items contained in the sequences of colors, pictures, and words had to be spoken to the subject. Presumably under this condition, the materials should be processed in identical codes, although it might be argued that categorical constraints were differentially involved.

Actually, it was the retarded subjects who displayed the most even performance across the various types of material presented in the auditory modality (see Table 1). Excluding the association spans, the retardates showed no significant threshold difference for materials under auditory presentation. There was more variability evident across thresholds in the performance of the second graders and particularly the sixth graders under auditory presentation. This finding suggests that the retardates may have relied much more heavily on an acoustic storage system and/or direct read out from that system.

Although it was not feasible to perform a factor analysis on the correlational data, an inspection of the correlation matrices suggests that, in general, there is a strong relationship among the various span measures. The average correlation was highest for the sixth-grade children and lowest for the special-class children, but substantial in all cases. These findings are consistent with conclusions from earlier studies regarding the operation of a general memory span factor (Brener, 1940). Looking at Table 3, there is also an indication that the correlations are somewhat higher within modalities than between modalities.

The one task that explicitly seems to involve reading processes is the visual word span. Marked differences were found between normal and retarded subjects with respect to the correlation between visual word span and spans for other types of visual materials. For both groups of normal children, visual word spans were highly correlated with performance on the other visual tasks; but this was not the case for the retarded subjects. It should be emphasized that all subjects, including retardates, were able to identify the printed

words. It may tentatively be suggested that the processes involved in reading enter more heavily into sequential short-term memory of normal children than of retarded children.

REFERENCES

Brener, R. An experimental investigation of memory span. *Journal of Experimental Psychology*, 1940, **26**, 467–482.

Cavanagh, J. P. Relation between the immediate memory span and the memory search rate. *Psychological Review*, 1972, **79**, 525–530.

Craik, F. I. M. Modality effects in short-term storage. *Journal of Verbal Learning and Verbal Behavior*, 1969, **8**, 658–664.

Ellis, N. R. Memory processes in retardates and normals. In N. R. Ellis (Ed.), *International review of research in mental retardation*. Vol. 4. New York: Academic Press, 1970.

Hawkins, W. F., & Baumeister, A. A. The effect of retroactive inhibition upon the digit span performance of normals and retardates. *American Journal of Mental Deficiency*, 1965, **69**, 871–876.

Jensen, A. R. A theory of primary and secondary familial mental retardation. In N. R. Ellis (Ed.), *International review of research in mental retardation*. Vol. 4. New York: Academic Press, 1970.

Korst, J. W., & Irwin, O. C. Immediate memory span of mentally retarded children. *The Cerebal Palsy Journal*, 1968, **29**, 10–11.

Laughery, K. R., & Pinkus, A. L. Short-term memory: Effects of acoustic similarity, presentation rate, and presentation mode. *Psychonomic Science*, 1966, **6**, 285–286.

MacMillan, D. L. Effect of input organization and recall of digits by EMR children. *American Journal of Mental Deficiency*, 1970, **74**, 692–696.

Madigan, S. A. Modality and recall order interactions in short-term memory for serial order. *Journal of Experimental Psychology*, 1971, **87**, 295–297.

Scott, K. G., & Scott, M. S. Research and theory in short-term memory. In N. R. Ellis (Ed.), *International review of research in mental retardation*. Vol. 3, New York: Academic Press, 1968.

Spitz, H. H. The channel capacity of educable mental retardates. In D. K. Rough (Ed.), *The experimental psychology of mental retardation*. Chicago: Aldine, 1973.

(Received January 29, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 895-902

READING GRADE LEVEL ADJUSTMENT AND AUDITORY SUPPLEMENTATION AS TECHNIQUES FOR INCREASING TEXTUAL COMPREHENSIBILITY[1]

ARTHUR I. SIEGEL[2] AND MARTIN R. LAUTMAN

Applied Psychological Services, Wayne, Pennsylvania

JAMES R. BURKETT

Air Force Human Resources Laboratory, Lowry Air Force Base, Colorado

Three experiments were performed to investigate situations in which lowering the reading grade level of textual material or providing tape-recorded auditory supplementation to the reading material would provide maximum comprehensibility gain. For typical Air Force technical training materials (study guides) and for an on-the-job look up manual, no gain was evidenced as the result of lowering the reading grade level or of auditory supplementation. However, for home study course materials, lowering the reading grade level resulted in increased criterion test scores.

Both during training and on the job, a high dependence is typically placed on written language for communicating information to technical personnel. Accordingly, materials that are written at a level which surpasses the literacy skill of technical personnel can be expected to mitigate on-the-job performance. Caylor, Sticht, Fox, & Ford (1972), for example, found that the reading grade level of military publications in the Army occupations they investigated exceeded the average reading ability of the men by from one to six years. Madden and Tupes (1966) documented the number of low-reading-ability airmen in the Air Force and Mowry, Webb, & Garvin (1955) showed that for Navy trainees, much of the required written technical materials was too difficult. Sticht and Kern (1971) found that, especially for low-aptitude personnel, the greater the gap between the literacy skills of the men and the literacy requirements of printed technical material, the less their tendency to use that material.

Since technical personnel employ many types of textual materials under varying conditions, clarification of the circumstances under which lowering the reading grade level of textual materials yields maximum benefit would seem important. This issue was the first concern of the present work. The second concern was the investigation of the effects of auditory supplementation (listening to tapes while reading) on comprehensibility. It seems that multimodal presentation could possess benefits because it would (*a*) serve to refocus the reader's attention in the case of distraction, (*b*) help the person who can comprehend the spoken language but who possesses a reading disability, and (*c*) serve as a redundant and noninterfering information source. There is conflicting evidence regarding the multimodal presentation issue. Long (1970),

[1] This research was completed under Contract Number F41609-72-C-0011 between the U.S. Air Force and Applied Psychological Services, Inc. Our indebtedness to both the various members of the administrative and teaching staffs at the Air Force materiel facilities specialist course of the Logistics School, Lowry Air Force Base, Colorado, as well as to Allan R. Williams, of the Applied Psychological Services, Inc., Wayne, Pennsylvania, staff who assisted in the organization and data collection of this study, is acknowledged.

[2] Requests for reprints should be sent to Arthur I. Siegel, Applied Psychological Services, Science Center, Wayne, Pennsylvania 19087.

Virag (1971), and Van Mondfrans and Travers (1962) have presented negative evidence while Severin (1967) and Singer (1970) have reported positive results. It is possible that the different types of materials and the varying conditions of use in these studies contributed to the discrepant findings.

Specifically, the hypothesis underlying this set of studies was that materials modification (either through lowering the reading grade level of textual materials or through multimodal presentation) would produce differential comprehensibility effects depending on the type of materials, the purpose of the materials, and the conditions of use. To this end, three types of textual materials (in-course study materials, home study materials, and a reference manual used on the job) were investigated within the Air Force materiel facilities occupational specialty. In all, three experiments were performed—one relative to each type of learning material and comprehensibility context.

Materials Development

All of the experiments employed written materials used in the Air Force materiel facilities (supply) specialty. The tasks performed by these specialists involve base supply ordering, inventory control, and distribution. The stimulus materials used in Experiment 1 consisted of a course study guide used during two intermediate weeks (three and four) of a formal, intensive, five-week materiel facilities specialist training course. The stimulus materials used in Experiment 2 consisted of career development course guides used as a home study manual by airmen who do not attend the formal training course but are seeking promotion within this supply specialty. The sections of the home study manual contained, basically, the same content as that covered in the formal training course. A technical manual (AFM 67-1, 1970) employed on the job as a reference manual by materiel facilities specialists was used in Experiment 3. At least one treatment group in each experiment received a simplified version of the appropriate stimulus materials. To develop the simplified materials, guidelines were followed which have been suggested for this purpose in the readability literature (Flesch, 1951; Klare, 1963). These were as follows: (*a*) using shorter words, (*b*) using shorter sentences, (*c*) avoiding "complicated" sentence constructions, (*d*) "personalizing" the material by the use of personal words and sentences, (*e*) using the active instead of the passive voice, and (*f*) presenting information in a stepwise manner.

For example, an unmodified original passage read as follows:

> The Air Force uses two basic methods—prepost and postpost—for issuing supplies. A prepost issue is processed through the computer before the property is actually selected from storage. A postpost issue is made before the issue request is processed through or into the computer. A simple way to determine which method is being used is that prepost documents are machine printed and postpost documents are normally handscribed.

The same passage simplified read as follows:

> The Air Force uses two methods for issuing supplies. These are prepost and postpost. A prepost issue is processed by the computer before the supplies are taken from storage. A postpost issue is processed by the computer after the supplies are taken from storage. You can easily tell which method is being used because prepost order documents are printed by machine. Postpost order documents are usually written by hand.

The simplified written revisions of the course study guide and home study manual materials and the unmodified originals were recorded on magnetic tape in cassette form. These tapes followed the written text verbatim and were identified for simultaneous use by subjects in certain treatment groups as they were reading the textual materials.

Materials Verification

The intent of the materials modification was to lower the reading grade level of the textual materials so that it would be within the reading grade level of the person using the materials for learning purposes. Such a modification would allow an evaluation of the effects of the modification on the ability of the different materials to transfer information. The automated readability index (ARI; Smith & Senter, 1967) and FORCAST (Caylor et al., 1972) formulas were first used for this purpose. These formulas

were selected rather than the more typically employed Dale-Chall (1948) and Flesch (1951) formulas since they were validated on military samples using technical materials. The ARI technique for estimating readability is based on word difficulty and sentence length. The FORCAST method is based on the number of one-syllable words. Both techniques yield a score which can be directly converted to a reading grade level for the textual material under consideration. Approximately 50% of the stimulus reading material, randomly selected from the modified and original course study guide, home study manual, and reference manual materials was analyzed using the two procedures. Table 1 shows the results of the analyses for the different stimulus materials.

The results indicated that the simplification procedures achieved the desired results, with the difference between the original and simplified versions showing up to a somewhat greater degree in the ARI-determined reading grade levels. The mean reading grade level of airmen in the Air Force supply field is 10.2. The ARI results suggested the original materials to be above this level and the modified materials to be considerably below this level. The FORCAST results indicated the modified materials to more closely match, but not to meet, the reading grade level of supply field personnel.

Convergent verification of the simplification procedure was also accomplished by using the cloze technique (Taylor, 1953). Eight randomly selected 150-word sample paragraphs from the original and the simplified course study guide and from the home study manual materials, matched for content, were administered to 39 airmen trainees in the first two weeks of the materiel facilities specialist training course. Each subject received both simplified and original materials (of different content) in a random order. Thus, each subject received eight course study guide and two home study manual paragraphs. The actual order of presentation of the passages to each subject was randomly determined. Following a standard cloze test procedure, every fifth word from a randomly determined starting point in each paragraph was deleted. Mean

TABLE 1

MEAN READING GRADE LEVEL AS MEASURED BY ARI AND FORCAST METHODS FOR ORIGINAL AND SIMPLIFIED MATERIALS

Material type	ARI		FORCAST	
	Original	Simplified	Original	Simplified
Study guides	11.53	6.80	11.75	10.02
HSM	12.74	7.40	11.95	11.20
AFM 67-1	14.57	8.32	12.53	11.09

Note. Abbreviations: ARI = automated readability index, HSM = home study manuals and AFM 67-1 = Air Force Manual (67-1).

cloze scores (number of deleted words correctly identified) for each trainee on the simplified and original versions were separately compared for the course study guide and the home study manual passages. In both the course study guide and home study manual passages, the simplified materials were found to be easier to read ($t = 4.00$, $df = 38$, $p < .001$ and $t = 2.72$, $df = 38$, $p < .01$, respectively).

Criterion Development

In order to test the comprehensibility of the original and the simplified textual materials, a set of criterion tests was developed using the multiple-choice format. To develop the course study guide and home study manual associated tests, an item pool was first assembled. Sixty multiple-choice items were written to cover the content of the course study guide for Weeks 3 and 4 (30 items per week), and 40 items were written covering the selected chapters of the home study manual. The number of test items developed was proportional to the amount of material covered in each section of the stimulus materials. The course study guide and home study manual test items were combined into a 100-item preliminary test which was administered to 63 materiel facilities specialist trainees who had completed Week 4 of the formal training course. Two preliminary test forms were constructed with the items presented in a reverse order on Form B. Approximately half of the subjects received each form. Item–test correlations and difficulty indices

were calculated for each item (Fan, 1952a, 1952b). An item–test correlation of $r \geq .21$ ($p \leq .10$) was accepted as a cutoff point for item inclusion in the final criterion test forms. Fifteen additional test items for each block were randomly selected from each of the criterion tests forms currently employed by course personnel. These items had been previously developed item analytically. The final criterion tests included 30 test items for each course week and 30 for the home study manual.

The multiple-choice format was also employed for the reference manual criterion test. The items for this test were not item analytically developed. They were selected, however, to represent a systematic sample of the paragraphs in the included sections for the reference manual.

Reading Grade Level Determination

Madden and Tupes (1966) provided equations for predicting the reading grade levels of Air Force enlistees from aptitude test scores obtained from the Airman Qualifying Examination (AQE). It is possible that the sample to which the formula would be applied in the present study would possess different characteristics from that on which the formulas were originally developed. To measure this possibility, a sample of 200 materiel facilities specialist trainees enrolled in the formal training course during the fall of 1971 was randomly selected. Application of the Madden and Tupes formula for administrative career fields to this new sample found the mean reading grade level to be 10.2—the identical value reported by Madden and Tupes for 1961–1962 data. The distributions of scores also appeared to be similar, and it was concluded that the equation was stable and could serve as a predictor of reading grade level for the subjects in this study.

Experiment 1

This experiment was designed to investigate the gain, if any, from lowering the reading grade level of study materials used as a classroom learning supplement and the learning advantages from supplementing the materials with a tape-recorded presentation which was listened to while the materials were read.

Method

Subjects

Sixty-six materiel facilities specialist trainees in the Logistics School at Lowry Air Force Base were selected to serve as subjects. All subjects were about to begin Week 3 of their five-week training course. It was necessary for each subject to complete successfully Week 3 prior to advancing to Week 4. Five subjects who failed the Week 3 criterion tests were forced to repeat Week 3. It was possible to place four of these subjects in classes receiving the same treatment, and they continued in the experiment. Each passed the Week 3 criterion test after the repetition and were advanced to Week 4. The result was a total of 65 subjects.

Experimental Design

Four treatment groups were formed with classes of subjects randomly assigned to a treatment group. A control group used as stimulus material the original, unsimplified written study guides; one experimental group used the original written study guides but, in addition, had the tape recordings of the original lessons. Subjects in this treatment group were instructed to listen to these tapes while simultaneously reading the required materials. The third group employed the simplified study guides, and the fourth group used both the simplified written materials as well as tape recordings of the simplified lessons. For analytic purposes, all subjects were postexperimentally divided into high- and low-reading-ability groups (split at a reading grade level of 10) by using the Madden and Tupes reading grade level determination formula for administrative career fields in the Air Force.

Procedure

All subjects, after completing Week 2 of the training course, were read instructions indicating the nature of the study in which they were about to participate. They were told that they had been chosen to take part in a program designed to compare various methods of presenting instructional materials and that they, personally, could benefit in the future from the results. For the treatment groups using auditory supplementation, a demonstration of the use of a Honeywell playback device was given. These subjects were additionally instructed to listen to the appropriate tape and to read simultaneously their study guides at least once, although they could read and listen to them as often as they wished. The use of the playback device was restricted to weekday nonclass hours. Monitoring by course personnel ensured the proper use of the auditory materials.

Each treatment group was composed of two classes. Members of each class progressed together

from Week 3 to Week 4, except for the failures who repeated Week 3 before progressing to Week 4. Written questionnaires (the questions were also read to the subjects) were administered on the final day of each block prior to the criterion test session. The questionnaires included two items which inquired into the difficulty and likability, respectively, of the written course study guide using five-point Likert-type items. Data for only the questionnaire administered at the end of Week 4 were analyzed to allow for maximum subject familiarity with the stimulus materials.

Results

The criterion test scores for Weeks 3 and 4 were summed for each subject independently for both the criterion test items developed specifically for this study using item-analytic procedures as well as for those randomly selected from the test questions normally used in the course. No statistically significant differences were found between the two sets of criterion scores (those developed here and those previously developed) in a repeated measures analysis of variance ($F = 1.62, df = 1/64, p > .05$), and the two sets of scores were pooled in the subsequent analysis. An unweighted means analysis of variance was performed on the criterion test data to determine the effects of experimental materials and reading grade level. A significant main effect for reading grade level was observed with the better readers scoring higher on the criterion tests ($F = 10.92$, $df = 1/47, p < .01$). No statistically significant experimental materials main effect ($F < 1, df = 3/47, p > .05$) or interaction effect ($F = 1.05, df = 3/47, p > .05$) was observed.

Two 2 × 2 unweighted means analyses were performed on the questionnaire items. The analyses compared the difficulty and likability, respectively, of the original and modified course study guide materials used in Weeks 3 and 4 with those used in Weeks 1 and 2 for the high- and low-reading-grade-level subjects. No statistically significant differences were observed for reading grade level in either analysis ($F < 1, df = 1/61$, $p > .05$ for difficulty; $F = 3.15, df = 1/61$, $p > .05$, for likability) or for experimental materials difficulty ($F < 1, df = 1/61$, $p > .05$). A statistically significant effect for experimental material likability was observed with the modified materials judged more likable ($F = 6.16, df = 1/61, p < .05$).

Experiment 2

Experiment 2 paralleled Experiment 1 in purpose but employed the home study manual materials.

Method

Subjects

Forty-nine Air Force trainees at Lowry Air Force Base awaiting assignment to schools for specialty training served as subjects.

Experimental Design

Three treatment groups were formed. The control group used the original, unmodified home study manual. One experimental group used the simplified home study manual and a second experimental group used the simplified home study manual with auditory supplementation. For analytic purposes, all subjects were postexperimentally divided into high- and low-readability groups (split at a reading grade level of 10).

Procedure

On arrival at the test center, subjects were randomly assigned to one of the three experimental treatments. Each treatment was administered in a separate room. The instructions given to the subjects were similar to those used in Experiment 1, noting that the purpose of the study was to compare various reading materials. The subjects were also told that after studying their materials for two hours (an amount of time judged by the course personnel to be more than adequate for even a slow reader), they would take a "very hard test" to measure how much they had learned. As before, the subjects in the auditory supplementation condition were told to listen and read simultaneously. Subjects were allowed to take breaks whenever they wished. However, they were repeatedly advised of the difficulty of the forthcoming test and told to study and learn as much as they could. After the study period, the home study manual materials were collected and the tapes were turned off. The 30-item multiple-choice test, developed as previously described, was then administered.

Results

An unweighted means analysis of variance was performed on the criterion test data to determine the effects of experimental materials and reading grade level. A statistically significant main effect for experimental materials ($F = 3.31, df = 2/43$, $p < .05$) was the only statistically signifi-

cant effect observed. Application of a Neuman-Keuls test indicated that the treatment group which received the simplified home study manual materials scored significantly higher than the group which received the original materials ($p < .05$). The group working with the simplified materials and with auditory supplementation scored intermediate between the other two and was not different from either of them at a statistically significant level. The high-reading-ability subjects received somewhat higher criterion scores than the low-reading-ability subjects. However, neither this difference ($F < 1$, $df = 1/43$, $p > .05$) nor the interaction effect ($F < 1$, $df = 2/43$, $p > .05$) was statistically significant.

Experiment 3

Experiment 3 paralleled the prior two experiments in purpose but employed the reference manual as the reading material.

Method

Subjects

Thirty-five airmen awaiting assignment to schools for specialty training at Lowry Air Force Base served as subjects. All subjects had also participated in Experiment 2.

Experimental Design

Two experimental groups were formed. One group used the original, unmodified reference manual while the second group used the simplified version. As in Experiments 1 and 2, for analytic purposes, all subjects were postexperimentally divided into high- and low-reading-ability groups (split at a reading grade level of 10).

Procedure

Experiment 3 was administered one day after Experiment 2. The subjects were regrouped according to their treatment groups in Experiment 2 and then half from each group was assigned to each of the two current experimental conditions. Each experimental condition was administered in a separate classroom. The subjects were told that they would be allowed 20 minutes to become familiar with the organization of the reference manual and that at the end of that time they would be given a 30-item, timed, multiple-choice test in which they were to answer correctly as many items as they could during the 20 minutes allotted. They were also told that they would be expected to look up answers during the test, but they were cautioned not to spend too much time on any one question, since less than one half of a minute was allotted for each question. Subjects were also told to guess on all questions about which they were unsure and to answer all questions. This procedure parallels the method of use of the reference manual in operational contexts.

Results

A 2 × 2 unweighted means analysis of variance was performed on the criterion test data. The only statistically significant main or interaction effect observed was for reading grade level ($F = 5.51$, $df = 1/33$, $p < .05$).

Discussion

The results of the three experiments here described provide insight into two major issues: (*a*) delineation of when a mismatch between the literacy skills of technical personnel and literacy requirements can be expected to result in less than effective performance and (*b*) evaluation of methods for minimizing such deleterious effects.

The results of Experiment 1 indicated a statistically significant main effect for reading grade level. This finding was not unexpected because of the known high correlation (Madden & Tupes, 1966) between Airmen Qualifying Examination scores (used here to predict reading grade level) and reading ability. Little if any gain, however, was realized from simplification of the textual materials or from the auditory supplementation. This seemingly surprising result might have been anticipated. First, the study guides, in the present context, were essentially a lecture and classroom supplementation. In the formal training situation involved here, repeated presentations of the course content took place through lecture, workbooks, programmed instruction handouts, remedial sessions, and the study guides themselves. Accordingly, if the student did not learn the subject matter content through one medium, he could master it through another. With multiple learning and review presentations, neither the reading grade level of the written material nor the auditory supplementation was overriding. A statistically significant difference in likability, however, was observed with the subjects showing a preference for the simplified materials.

Experiment 2, which involved learning through the home study manual materials only, did not allow for the "information redundancy" of Experiment 1. Although the high-reading-ability subjects received somewhat higher criterion scores than the low-reading-ability subjects, the difference was not statistically significant. It is possible that the short study time and difficult learning task did not permit high reading ability to manifest itself to advantage. More importantly, a statistically significant difference among the treatment groups was found in Experiment 2. The group which received the simplified home study manual materials scored higher than the group which received the original materials. It is clear, then, that when the luxury of "information redundancy," which might be present in a formal training course, is absent, rewriting of the course materials at a lower reading grade level can be expected to lead to more effective performance. With reference to the criterion involved, however, there is only little in the present data to support the contention favoring the use of simultaneous auditory supplementation.

The present findings differ from those of Sellman (1970), who also used home study manual materials to investigate the effects of lowered reading grade level and audio supplementation on criterion test performance as a function of mental ability category. His results showed a simplified, materials-plus-tapes group to score higher than groups using only simplified materials or conventional materials. No differences were found between the latter two groups. Mental ability was also found by him to be statistically significant, with subjects in the higher mental ability categories performing better. There are, however, several differences between Sellman's procedure and that employed here. First, no time limit was imposed on his subjects, with the consequence that the simplified-plus-tapes group spent a statistically significant greater amount of time (as reported by Sellman) with the materials than either of his other two groups. It would thus appear that time for study was confounded with the experimental treatments. Second, his "tapes" condition involved "reiterations of topics already discussed." This procedure results in enforced practice by the "tapes" group rather than the simple, simultaneous audio presentation as employed in this study. Finally, Sellman administered the same pre- and posttest to his subjects. His results on mental ability categories thus might be somewhat artifactual since he may have been measuring "ability to remember the questions."

Experiment 3 involved a technical manual used for reference purposes on the job. As in Experiment 1, the better readers scored higher on the criterion test. No significant treatment effects were found in Experiment 3 and, accordingly, it seems that with a "look-up" type of task, reading grade level simplification, as here employed, exerts little practical effect. Possibly, in a look-up-type multiple-choice test, subjects merely have to recognize answers and not, necessarily, comprehend the materials. It is also possible that the material was too difficult and/or the time period too short for a treatment effect to appear. Nevertheless, simplification of technical materials of this sort might be expected to have only marginal effects on performance. Higher reading grade level, on the other hand, can be expected to read faster and, hence, complete more items correctly.

The fact that the ARI and FORCAST formulas showed high agreement in their reading-grade-level determination for the original materials but not for the simplified materials (although both formulas agreed in direction of the effect), is probably indicative of their differential sensitivity to the simplifying procedure utilized. The rewriting procedure clearly favored the short sentence – small word approach of the ARI. It has been noted, however, that readability formulas may not apply when the material being measured has been intentionally simplified (Klare, 1963). Independent verification of the simplifying procedure employed here by using the cloze procedure would thus seem to be critical in any future efforts using the paradigm of this study.

The issue of using grade levels both as measures of reading-level achievement as well as of prose difficulty is one that should be considered. Clearly, although a common

metric is being used, the constructs are not congruent. Moreover, the reading grade level of the trainees was estimated here by using an aptitude test (Airmen Qualifying Examination) score. Accordingly, the equation of passage reading difficulty and subject reading ability by virtue of equal reading grade level assignment is something less than a straightforward analogy. In the absence of a common metric able to perform this task, however, the procedure followed here would seem adequate.

Finally, the traditional criterion test development technique followed here favors the inclusion of items that maximize individual differences. Thus, the failure to find treatment effects on Experiments 1 and 3 could have possibly been the result of the insensitivity of the criterion test to such an effect. On the other hand, if one adopts this view, the statistically significant treatment effect observed in Experiment 2 is probably a very stable finding.

In summary, it would seem that the maximum gain will accrue by writing textual materials which are used in situations in which there is little opportunity for redundant information, at a reading grade level which is concordant with the ability of the reader. This situation is typified by the home study manual experiment described here (Experiment 2). Lesser gain, if any, is indicated by emphasis on modification of materials employed as adjuncts to other materials or for on-the-job reference purposes.

REFERENCES

AFM 67-1, *Supply manual*. Washington, D.C.: Department of the Air Force, 1970.

Caylor, J. S., Sticht, T. G., Fox, L. C., & Ford, J. P. *Methodologies for determining reading requirements of military occupational specialities*. (Draft Technical Report) Presidio of Monterey, Calif.: Human Resources Research Organization, January 1972.

Dale, E., & Chall, J. S. A formula for predicting readability. *Educational Research Bulletin*, 1948, **27**, 11–20.

Fan, C. T. *Item analysis table*. Princeton, N.J.: Educational Testing Service, 1952. (a)

Fan, C. T. *Note on construction of an item analysis table for the high-low 27 per-cent group method*. Princeton, N.J.: Educational Testing Service, 1952. (b)

Flesch, R. *How to test readability*. New York: Harper, 1951.

Klare, G. R. *The measurement of readability*. Ames: Iowa State University Press, 1963.

Long, G. R. A study on the preparation of an audio-tutorial minicourse. *Dissertation Abstracts International*, 1970, **31**, 1690A–1691A.

Madden, H. L., & Tupes, E. C. *Estimating reading ability level from the AQE General Aptitude Index*. (Tech. Rep. PRL-TR-66-1) Lackland Air Force Base, Texas: Personnel Research Laboratory, Aerospace Medical Division, February 1966.

Mowry, H. W., Webb, W. B., & Garvin E. A. *Readability of instructional materials and methods for analyzing*. (Tech. Rep. 2) Memphis, Tenn.: Naval Air Technical Training Command, 1955.

Phillips, M. Learning materials and their implementation. *Review of Educational Research*, 1966, **36**, 373–379.

Sellman, W. S. *Effectiveness of experimental training materials for low aptitude airmen*. (Tech. Rep. AFHRL-TR-70-16) Lowry Air Force Base, Colo.: Air Force Human Relations Laboratory, Technical Training Division, June 1970.

Severin, W. The effectiveness of relevant pictures in multiple channel communications. *AV Communication Review*, 1967, **15**, 386–401.

Singer, C. R. Eye movements, reading comprehensions, and vocabulary as effected by varying visual and auditory modalities in college students. *Dissertation Abstracts International*, 1970, **31**, 164A.

Smith, E. A., & Senter, R. J. *Automated readability index*. (Tech. Rep. AMRL-TR-66-220) Wright Patterson Air Force Base, Ohio: Aerospace Medical Division, November 1967.

Sticht, T. G., & Kern, R. P. *Project REALISTIC: Determining literacy demands of jobs*. (Tech. Rep. 3-71) Presidio of Monterey, Calif.: Human Resources Research Organization, April 1971.

Tavlor, W. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 1953, **30**, 415–433.

Van Mondfrans, A., & Travers, R. Paired-associate learning within and across sense modalities and involving simultaneous and sequential presentations. *American Educational Research Journal*, 1962, **2**, 89–99.

Virag. W. The effectiveness of three instructional modes employed to transmit content to students with different aptitude patterns. *Dissertation Abstracts International*, 1971, **31**, 5520A.

(Received January 29, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 903-910

INFORMATION SEEKING FOLLOWING THE CONFIRMATION OR CONTRADICTION OF BELIEFS[1]

CHARLES B. SCHULTZ[2]

Trinity College

The effect of contradiction and confirmation on information seeking and acquisition was investigated. Seventy-five undergraduates selected and examined the more interesting slide in a pair of slides. The critical slide pairs were comprised of one member containing belief-congruent information and another containing belief-discrepant information. Before the slides were viewed, evidence was presented which either contradicted existing beliefs (contradiction), supported existing beliefs (confirmation), or contradicted and supported existing beliefs (doubt). Contradiction subjects examined the slides longer and acquired more information than confirmation subjects. Ratings of interest, selection of slides, examination of slides, and acquisition of information were in a belief-congruent direction for the confirmation condition and a belief-discrepant direction for the contradiction condition.

Learners are often confronted with arguments or evidence which conflict with beliefs they hold. Such discrepant information is thought to have at least two effects on further information seeking and acquisition: (*a*) According to one theory of curiosity (Berlyne, 1960, 1965), support for a discrepant position sharpens conceptual conflict by strengthening a subordinate belief to a point where it competes with the dominant belief. As a result of the heightened conflict and the curiosity which accompanies it, information seeking is intensified and, eventually, reinforced when the search is brought to a satisfactory conclusion. (*b*) According to dissonance theory (Festinger, 1957, 1964), information regarding a discrepant position in particular is not sought and, in fact, is avoided because it is more likely to arouse or increase dissonance than to reduce it.

In regard to the first of these theories, Berlyne (1962) manipulated conceptual conflict by varying the number and the equiprobability of response alternatives (i.e., uncertainty). Both factors contributed to curiosity. However, curiosity was measured in this study and in others (e.g., Heslin, Blake, & Rotton, 1972) by self-reports of intentions to seek additional information rather than by search behavior. A direct relationship between uncertainty and information search has been demonstrated on a variety of tasks ranging from complex verbal problems (Lanzetta, 1963) to concept learning (Long, 1965), word games, and picture identification (Lanzetta & Driscoll, 1968). Moreover, this relationship appears to be mediated by subjective uncertainty (i.e., the subject's estimate of uncertainty) which on the one hand, increases monotonically with uncertainty due to stimulus and response factors (Driscoll & Lanzetta, 1965), and, on the other, is directly related to information search (Driscoll, Tognoli, & Lanzetta, 1966).

These particular findings on the effect of uncertainty on search behavior are consistent with Berlyne's (1960, 1965) conflict theory of curiosity. Berlyne's (1965) theory also includes the assumption that information which reduces relatively large amounts of curiosity will be strongly reinforced. Thus, not only is information-seeking be-

[1] The research reported herein represents one portion of a doctoral thesis which was supported in part by the Advanced Research Projects Agency (ARPA Order 1269) through the Office of Naval Research under Contract ONR Nonr N0014-67-A-0385-0006.

The author is particularly indebted to Professor Francis J. DiVesta, thesis advisor, for his invaluable assistance and encouragement.

[2] Requests for reprints should be sent to Charles B. Schultz, Trinity College, Hartford, Connecticut 06106.

havior reinforced by the reduction of conceptual conflict and, therefore, more likely to occur on similar occasions, but the knowledge derived from the search is more likely to be retained than knowledge which is not required to reduce conceptual conflict. In this regard, Berlyne (1954) found that subjects recalled more information from statements when they were asked questions beforehand and when they regarded the statements as familiar. Berlyne inferred conceptual conflict from ratings of familiarity and from the effect of prequestions. A primary purpose of the present investigation was to examine the effect of directly inducing conceptual conflict by strengthening a subordinate response on both information seeking and the acquisition of knowledge.

Often information which is necessary for the reduction of conceptual conflict is discrepant with an individual's existing beliefs. However, in most cases, information is selectively sought in order to avoid exposure to a contradictory position (Festinger, 1957, 1964). At least one exception to this general rule occurs when the contradictory or discrepant information is thought to be of practical use (Canon, 1964; Freedman, 1965). Discrepant information also may be sought for its *intrinsic utility*, that is, for its usefulness in resolving the conceptual conflict itself and the ensuing curiosity (Rhine, 1967). Indeed, Festinger (1964) notes that "avoidance [of discrepant information] would be observed only under circumstances where other reasons for exposure, such as usefulness or curiosity, were absent [p. 96]."

In the present study, two forms of conceptual conflict were examined to determine under which conditions, if any, selective exposure is modified. In one condition, only discrepant evidence was presented to the subject (contradiction) before examining additional materials which contained both congruent and discrepant information; in the other condition, both congruent and discrepant evidence were presented (doubt). Since contradiction involves strengthening only the subordinate belief while doubt involves strengthening both subordinate and dominant beliefs, contradiction may direct the individual more explicitly toward the discrepant position as a potential source of information for the resolution of conceptual conflict than doubt. Indeed, discrepant information may be indispensable for the resolution of conflict in the contradiction condition, and, therefore, be of considerable intrinsic utility. Thus, "playing the devil's advocate" may be instructionally productive since it prods learners to seek information which disagrees with their existing beliefs and which could increase dissonance.

Two hypotheses were derived from the above rationale: (*a*) When conceptual conflict is aroused by discrepant evidence presented by itself or in combination with supporting evidence, an individual seeks and acquires more information on the topic in general than he otherwise would. (*b*) Particularly under conditions of contradiction, the individual seeks and acquires more information on the discrepant position itself than when his existing beliefs are confirmed.

Method

Subjects

The subjects were 75 volunteers from an undergraduate course in educational psychology at the Pennsylvania State University who earned standard score points toward their grade for participating in the experiment. There were 15 subjects randomly assigned to each treatment. An additional 15 were assigned to the doubt group because data from that cell were analyzed for other experimental purposes reported elsewhere (Schultz, 1974).

Stimulus Materials

The principal experimental materials were a modified version of the Festinger and Carlsmith (1959) study of the cognitive effects of forced compliance. At least two advantages were obtained by using the Festinger-Carlsmith study as experimental materials: Common sense explanations of attitude change can be made according to two conflicting theories (dissonance vs. reinforcement) and the reinforcement position is almost unanimously invoked as an explanation of attitude change by relatively naive observers. All 75 subjects included in the analyses selected the reinforcement position. Five other subjects were excluded because they chose the dissonance position. Thus, information about reinforcement theory was congruent with the subject's belief, while information about dissonance theory was discrepant.

The description of the Festinger-Carlsmith study was broken down into small units of a paragraph or several lines which were presented via 2 × 2 inch slides. Two slides were rear projected simultaneously from two carousel projectors onto

a translucent screen. In all, there were 30 slide pairs which were organized into three segments. The 16 pairs of slides in the second segment were particularly important because they contained definitions, assumptions, rationales, and predictions based on the congruent and discrepant positions. Each pair consisted of one congruent and one discrepant slide. The congruent slide was balanced in length and grammatical form to match the discrepant member on the same topic. As an illustration, the prediction based on dissonance theory appeared on the screen next to the prediction based on reinforcement theory. The slides were titled to identify the content and to indicate at a glance which slide in the pair was congruent and which was discrepant. For the sake of simplicity, reinforcement theory was labeled the "law of reward" and dissonance theory the "theory of conflict." The congruent and discrepant slides were randomly assigned to the two projectors to prevent either type from appearing consistently on one side of the screen.

The first segment of slides (seven slide pairs) described Festinger and Carlsmith's experimental procedures. The third segment (seven slide pairs) described the experimental results. The two slides in each pair in both of these segments were identical, a procedure adopted to permit the use of the same slide-changing routine throughout the experiment.

Procedure

The subject was given a brief overview of the tasks he was to perform including practice in the slide-changing routine. This routine consisted of pressing two control buttons simultaneously, projecting a slide pair on the screen. The subject decided which was the less interesting slide in the pair, turned it off, and examined the remaining slide. When he finished, that slide was turned off to reveal a blank screen which was the cue to begin another cycle by pressing both control buttons. The subject was free to set his own pace.

Because subjects were not told that they would be tested on the slide content, the task of examining the slides required a plausible explanation. Accordingly, it was explained that the experimenters were preparing instructional materials on the topic of attitude change and that the subject could help by identifying what he considered to be the most interesting slides. It was stressed that only his actual personal preferences would be useful.

At this point, a problem on attitude change was described to the subject who was given a choice between two alternative outcomes. The subject had to decide whether a large or a small reward was most effective in changing attitudes. Conceptual conflict was induced by presenting contrived results of an experiment on this topic. The subjects in the *contradiction* condition were given experimental evidence which disagreed with their selection of the larger payment. The subjects in the *confirmation* condition were given evidence which agreed with their selection of the larger payment. In the *doubt* condition, subjects were told that two experiments were conducted. They were shown the same bogus evidence given the confirmation subjects as well as the evidence given to the contradiction subjects. An absolute *control* group differed from the experimental groups only in that it received neither supporting nor contradictory evidence.

Measures

Each time the subject turned either of the slide projectors on or off, his response was registered by an event recorder. The measure of information seeking obtained from this procedure was *examination time*, that is, the time the subject spent examining the more interesting member of any slide pair after turning off the less interesting slide. *Reaction time* was the interval between turning on both projectors and turning off the less interesting slide in the pair. Information acquisition was determined by a 33-item multiple-choice test of the content on the slides.

Ratings of interest in "reward" and "conflict" slides provided one measure of selective exposure. In addition, the subject rated his interest in six "reprints from psychological journals." The titles of two of these articles were congruent with his beliefs, two were neutral, and two were discrepant. Ratings for each type of article were averaged to provide a measure of selective exposure.

For each of the 16 pairs of slides in the second segment, the subject had the opportunity to select congruent or discrepant information. The number of discrepant slides chosen (DSC) was tabulated. Since the subject may have selected the discrepant slides more often than the congruent slides, but spent less time examining them, a discrepant time/examination time ratio (D/E ratio) was computed to reflect the proportion of time spent examining discrepant information. Discrepant time is the examination time for only the discrepant slides.

The multiple-choice test was comprised of part scores with items which required knowledge of congruent information (nine items), discrepant information (seven items), and both congruent and discrepant information (seven items). Other items required knowledge of Festinger and Carlsmith's (1959) experimental procedures and results.

Results

In order to assess the extent to which the treatments had been induced, subjects rated the amount of conflict they perceived between slides in a pair on a 100-point scale at the end of the experiment. A single-factor analysis of variance was conducted on these scores. Although the differences fell short of significance ($F = 2.02$, $df = 3/71$, $p > .05$), the direction of the means was

TABLE 1
MEANS AND STANDARD DEVIATIONS FOR REACTION TIME, EXAMINATION TIMES, AND ACQUISITION OF KNOWLEDGE

Measure	Experimental condition			
	Contradiction (n = 15)	Doubt (n = 30)	Control (n = 15)	Confirmation (n = 15)
Reaction time[a]				
$\bar{X}$	155.73	235.25	165.47	148.80
SD	95.12	148.43	112.05	113.51
Examination time[a]				
$\bar{X}$	158.33	126.98	101.70	93.50
SD	55.88	70.21	66.67	35.01
Examination time[b]				
$\bar{X}$	97.17	89.32	81.97	78.20
SD	34.10	46.27	25.67	25.62
Test score[c]				
$\bar{X}$	18.67	20.47	17.40	14.93
SD	4.03	3.45	3.40	3.95

[a] For congruent and discrepant slides (in seconds).
[b] For Festinger-Carlsmith results (in seconds).
[c] Items correct.

consistent with expectations and with results on dependent measures reported below. Both the contradiction ($\bar{X}$ = 74.81) and doubt ($\bar{X}$ = 74.10) groups reported greater conflict than the control ($\bar{X}$ = 69.33) while the confirmation group ($\bar{X}$ = 58.47) reported less.

Findings relevant to the hypothesis that contradiction results in greater examination and recall than confirmation are summarized in Table 1. The analysis of variance of examination time devoted to slides which paired congruent and discrepant information (Segment 2 slides) was significant (F = 3.46, df = 3/71, $p < .025$). In order to test pairwise differences among the means, Dunn's multiple-comparison procedures were used in this and all other analyses (p = .05). Examination time for the contradiction condition was significantly longer than for the confirmation condition. The examination time for pairs of identical slides which depicted the results of the Festinger-Carlsmith experiment (Segment 3 slides) was similar to the examination times for Segment 2, but the differences in this case were not significant ($p > .05$). A similar analysis of variance revealed that the groups differed significantly in the amount of information acquired (F = 8.05, df = 3/71, $p < .001$). Both the doubt and the contradiction conditions had higher scores on the multiple-choice test than the confirmation condition according to Dunn's procedures, although they did not differ from each other. Finally, doubt subjects spent more time than those in other conditions deciding whether the congruent or discrepant member of each pair of slides in Segment 2 was more interesting; however, this difference was not reliable (F = 1.82, df = 3/71, $p > .05$).

Selective exposure was measured by the selection of articles for future reading which agreed (congruent) or disagreed (discrepant) with subject's existing beliefs, the interest reported in congruent or discrepant slides, the discrepant slide choice (DSC), the time spent examining discrepant slides (D/E ratio), and finally, the acquisition of information about the congruent or discrepant positions. The findings obtained on each dependent measure will be examined in turn.

The means displayed in Table 2 summarize the ratings of interest in reading articles which are congruent or discrepant with a belief the subject holds or which are unrelated to that belief (i.e., neutral articles). Since neutral articles neither supported nor contradicted the subject's existing beliefs, there was no reason to expect differences in the ratings of these articles across treatments. A single-factor analysis of variance yielded no significant differences among the treatment groups. However, a similar analysis of the selection of congruent articles was significant (F = 3.65, df = 3/71, $p < .025$). Congruent articles were rated more interesting by confirmation subjects than by subjects in the contradiction condition. Conversely, discrepant articles tended to be rated more interesting by subjects in the contradiction and doubt conditions than by subjects in the confirmation condition, although these differences were less reliable than those for congruent articles (F = 1.74, df = 3/71, $p > .05$).

These findings imply a Type of Article × Level of Conceptual Conflict interaction. In order to test this assumption, the ratings of congruent and discrepant articles were treated as repeated measures in a 2 × 2

TABLE 2
MEANS FOR MEASURES OF SELECTIVE EXPOSURE

Measure	Experimental condition			
	Contradiction (n = 15)	Doubt ((n = 30)	Confirmation (n = 15)	Control (n = 15)
Ratings of interest in articles (100-point scale)				
Congruent	61.07	68.40	77.33	74.67
Neutral	66.20	63.37	59.60	60.33
Discrepant	70.73	67.53	59.53	56.73
Ratings of interest in slides (100-point scale)				
Congruent	43.00	63.40	75.13	69.47
Discrepant	79.93	68.70	51.33	55.27
Choice and examination of slides				
Choice of discrepant slides	12.07	6.43	6.20	5.26
D/E ratio	.76	.40	.36	.34
Part scores on multiple-choice test (in percentages)				
Congruent information (n = 9)	45.13	62.63	60.73	48.07
Congruent and discrepant information (n = 7)	52.33	60.47	57.13	41.00
Discrepant information (n = 7)	67.60	64.80	37.07	41.93

mixed factorial analysis of variance design with two treatment levels (confirmation and contradiction) and two repeated measures (congruent and discrepant ratings). This analysis yielded a significant Treatment × Type of Article interaction (F = 7.48, df = 1/28, $p \leq .025$).

A single-factor analysis of interest reported in congruent slides yielded significant effects (F = 5.89, df = 3/71, $p < .005$) which were consistent with those regarding the selection of articles. Contradiction subjects rated the congruent slides as less interesting than subjects in each of the other three groups. A similar analysis of discrepant slides also yielded significant effects (F = 3.58, df = 3/71, $p \leq .025$) in which contradiction subjects rated discrepant slides higher than confirmation subjects. Since these findings also imply a Treatment × Type of Slide interaction, a 2 × 2 analysis of variance was conducted with ratings of congruent and discrepant slides treated as repeated measures. According to this analysis, there was a significant interaction effect (F = 14.61, df = 1/28, $p < .001$).

Single-factor analyses of variance yielded significant effects on both the number of discrepant slides chosen (F = 5.50, df = 3/71, $p < .01$) and the ratio of discrepant time to examination time (F = 5.28, df = 3/71, $p < .01$). The differences among the means were in the predicted direction and consistent with the findings based on interest scales (see Table 2). The greatest DSC and D/E ratios occurred in the contradiction condition, while the smallest DSC and D/E ratios occurred in the confirmation condition. Analyses of pairwise comparisons of both measures using Dunn's procedures indicated that all treatments differed from the contradiction treatment ($p < .05$) but not from each other.

The scores from three sections of the multiple-choice test of the experimental topic were converted to percentages to permit comparison (Table 2). According to a single-factor analysis of variance, there was a significant treatment effect on the set

of questions requiring both congruent and discrepant information ($F = 2.88$, $df = 3/71$, $p < .05$). Although scores of the doubt condition were higher than the other experimental treatments, Dunn's procedures revealed that the difference was significant only for the control group. Analysis of the set of items which measured retention of congruent information yielded a significant treatment effect ($F = 3.31$, $df = 3/71$, $p < .05$) in which confirmation subjects had higher scores than contradiction subjects. In contrast, the analysis of the items which measured retention of discrepant information yielded an F of 7.98 ($df = 3/71$, $p < .01$) in which confirmation and control subjects scored lower than contradiction and doubt subjects. Since these results imply a Treatments × Kind of Information Retained interaction, a repeated measures analysis of variance was conducted using the procedures described above. This analysis yielded a significant Treatment × Type of Item interaction ($F = 15.07$, $df = 1/28$, $p < .001$).

Discussion

According to the present findings, the presentation of evidence which contradicts existing beliefs produces greater interest in and examination of the experimental topic than would otherwise occur. Not only was more information sought when evidence supporting a subordinate belief was presented alone or paired with evidence supporting a dominant belief, but it was also retained. This finding, in particular, is consistent with Berlyne's (1960, 1965) theory of epistemic curiosity.

Exactly what accounts for the greater retention in the contradiction and doubt conditions is not clear from the present results. In regard to predecision information seeking, at least the effect of uncertainty is primarily reflected on indices of information processing (Driscoll & Lanzetta, 1964). Accordingly, the greater retention in the contradiction and doubt conditions may have been due to processes such as rehearsal or reorganization. In any event, the superior retention obtained by strengthening a subordinate response alternative extends findings by Berlyne (1954) and later by Frick and Cofer (1972) which demonstrated that curiosity generated by prequestions results in greater retention.

The preference for congruent information by subject in both the confirmation and control conditions, which did not differ from each other, clearly supports the selective exposure hypothesis derived from Festinger's (1957) theory of cognitive dissonance. However, the tendency of subjects in the contradiction condition to select and acquire discrepant information is difficult to account for by the theory of cognitive dissonance unless one assumes that contradictory evidence produced extreme dissonance. Typically, one would expect the presentation of evidence which contradicts subjects' existing beliefs to increase dissonance moderately, thereby resulting in a greater selection of congruent information than would occur without a dissonance increasing manipulation (Freedman & Sears, 1965).

The differences in selectivity between the confirmation and contradiction conditions may be due, in part, to the greater intrinsic utility the discrepant information held for the contradiction subjects. According to this hypothesis, discrepant information will be sought when knowledge of the existing belief is not sufficient to reduce curiosity. Thus, contradictory information is sought for its intrinsic utility, even at the risk of increasing dissonance and just as it is sought for its practical or extrinsic utility (Canon, 1964; Freedman, 1965).

This formulation, which appears to be supported be the present findings, may explain other instances in which discrepant information was not avoided, or it was sought. For example, in one study, subjects read heavily biased case reports of a criminal trial which favored either the defense or prosecution summation (Sears, 1965). The results indicated that subjects preferred information which was opposite both to the biased "factual" report they had read and to their own opinions. Since it is well known that there are both defense and prosecution sides to any legal issue, the mock trial experimental situation may have generated curiosity which required discrepant information for its reduction. Similar interpretations can be made of other mock trial studies in

which discrepant information was preferred or in which preference for congruent information was not found (Sears, 1966; Sears & Freedman, 1965).

To the extent that the present analysis is valid, intrinsic utility is a factor to be investigated in its own right as well as a variable to be controlled in selective exposure experiments. Since whatever creates dissonance may also arouse epistemic curiosity, the effects of the two variables may be difficult, if not impossible, to separate (Rhine, 1967). Nevertheless, in the present experiment, only the explicit contradiction of the subjects' beliefs generated the amount or type of conflict necessary to direct them toward discrepant information. Whatever conflict was created by simply justapositioning both congruent and discrepant information in the slide pairs of Segment 2 (i.e., the control condition) was not sufficient to prompt subjects to risk the increase in dissonance implied by the examination of discrepant information.

Differences in selectivity between the contradiction and doubt conditions are particularly interesting and are especially important because of their implication for education. Contradiction typically directs the subject's interest in and examination and retention of information toward the discrepant position at the expense of congruent information. The effects of doubt are not so myopic. On almost every measure doubt scores are relatively high. Moreover, they are consistently higher than confirmation scores on measures of preference for discrepant information and higher than contradiction scores on measures of preference for congruent information (Table 2). Only on the measures of choice of discrepant slides and time spent examining discrepant slides does the behavior of doubt subjects resemble that of confirmation subjects. This effect may be explained by the tendency of doubt subjects to spend more time looking at both slides in a pair before turning off the less interesting member.

Thus, doubt appears to produce a relatively balanced examination of congruent and discrepant views which is necessary for the study of controversial issues. In contrast, confirmation directs information seeking toward a congruent position in apparent accordance with dissonance theory, while contradiction directs information seeking toward a discrepant position, perhaps because of the intrinsic utility that information has acquired.

REFERENCES

Berlyne, D. E. An experimental study of human curiosity. *British Journal of Psychology*, 1954, **45**, 256–265.

Berlyne, D. E. *Conflict, arousal, and curiosity.* New York: McGraw-Hill, 1960.

Berlyne, D. E. Uncertainty and epistemic curiosity. *British Journal of Psychology*, 1962, **53**, 27–34.

Berlyne, D. E. *Structure and direction in thinking.* New York: Wiley, 1965.

Canon, L. K. Self-confidence and selective exposure to information. In L. Festinger (Ed.), *Conflict, decision, and dissonance.* London: Tavistock, 1964.

Driscoll, J. M., & Lanzetta, J. T. Effects of problem uncertainty and prior arousal on pre-decision information search. *Psychological Reports*, 1964, **14**, 975–988.

Driscoll, J. M., & Lanzetta, J. T. Effects of two sources of uncertainty in decision making. *Psychological Reports*, 1965, **17**, 635–648.

Driscoll, J. M., Tognoli, J., & Lanzetta, J. T. Choice conflict and subjective uncertainty in decision making. *Psychological Reports*, 1966, **18**, 417–432.

Festinger, L. *A theory of cognitive dissonance.* Stanford, Calif: Stanford University Press, 1957.

Festinger, L. *Conflict, decision, and dissonance.* London: Tavistock, 1964.

Festinger, L., & Carlsmith, J. M. Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, 1959, **58**, 203–210.

Freedman, J. L. Preference for dissonant information. *Journal of Personality and Social Psychology*, 1965, **2**, 287–289.

Freedman, J. L., & Sears, D. O. Selective exposure. In L. Berkowitz (Ed.), *Advances in experimental psychology.* Vol. 2, New York: Academic Press, 1965.

Frick, J. W., & Cofer, C. N. Berlyne's demonstration of epistemic curiosity: An experimental re-evaluation. *British Journal of Psychology*, 1972, **63**, 221–228.

Heslin, R., Blake, B., & Rotton, J. Information search as a function of stimulus uncertainty and the importance of the response. *Journal of Personality and Social Psychology*, 1972, **23**, 333–339.

Lanzetta, J. T. Information acquisition in decision making. In O. J. Harvey, *Motivation and social interaction.* New York: Ronald Press, 1963.

Lanzetta, J. T., & Driscoll, J. M. Effects of uncertainty and importance on information search

in decision making. *Journal of Personality and Social Psychology*, 1968, **10**, 479–486.

Long, B. H. Predecision search in concept formation: The effects of problem complexity. *The Psychological Record*, 1965, **15**, 197–202.

Rhine, R. J. Some problems in dissonance theory research on information selectivity. *Psychological Bulletin*, 1967, **68**, 21–28.

Schultz, C. B. The effect of confidence on selective exposure: An unresolved dilemma. *The Journal of Social Psychology*, 1974, **94**, 65–69.

Sears, D. O. Biased indoctrination and selectivity of exposure to new information. *Sociometry*, 1965, **28**, 363–376.

Sears, D. O. Opinion formation and information preferences in an adversary situation. *Journal of Experimental Social Psychology*, 1966, **2**, 120–142.

Sears, D. O., & Freedman, J. L. The effects of expected familiarity with arguments upon opinion change and selective exposure. *Journal of Personality and Social Psychology*, 1965, **2**, 420–426.

(Received February 18, 1974)

Manuscripts Accepted for Publication in the *Journal of Educational Psychology*

Grades Expected and Grades Received—Their Relationship to Students' Evaluations of Faculty Performance. W. Robert Kennedy (School of Medicine, Division of Research in Medical Education, Case Western Reserve University, 2119 Abington Road, Cleveland, Ohio 44106).

Differential Effects of Two Grading Systems on Student Performance. Reed G. Williams (School of Medicine, Southern Illinois University, P.O. Box 3926, Springfield, Illinois 62708), Martin J. Pollack, and Nancy A. Ferguson.

Effect of Information Feedback on Matrices Learning for Differing Socioenomic Levels. Ralph R. Turner (Department of Psychology, West Virginia University, Morgantown, West Virginia 26506).

The High School Transcript as a Set of "Nonreactive" Measures for Predicting College Success and Major Field. Roy D. Goldman (Department of Psychology, University of California, Riverside, California 92502) and Dennis W. Sexton.

Instructional Treatments, Presenter Types, and Learner Characteristics as Significant Variants in Instructional Television for Adults. Robert D. Brown (Department of Educational Psychology and Measurements, Teachers College, University of Nebraska, Lincoln, Nebraska 68508), Laurence A. Brown, and James E. Danielson.

Word Consciousness in Readers and Prereaders. Linnea C. Ehri (Department of Education, College of Letters and Sciences, University of California, Davis, California, 95616).

Peer Acceptance and School-Related Variables in An Integrated Junior High School. Donald E. Carter (Child Study Center, State University College at Buffalo, 1300 Elmwood Avenue, Buffalo, New York 14222), Susan L. DeTine, June Spero, and Forrest W. Benson.

Retention of Prose Materials as a Function of Postacquisition Testing. James F. Voss (Department of Psychology, Faculty of Arts and Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260) and Ronald E. La Porte.

Delayed Retention Effect in Natural Classroom Settings. Richard C. Anderson (University of Illinois College of Education, Department of Educational Psychology, Urbana, Illinois 61801) and John R. Surber.

Heat, Ethnic Differences and Creativity in the Negev Desert. Alvin H. Shapiro (Ontario Ministry of Health, Box 930, Postal Station P., Thunder Bay Ontario, Canada P7B 5G4).

Personal Characteristics and Teaching Effectiveness of College Faculty. Barbara R. Sherman (Academic Affairs, Western Michigan University, Kalamazoo, Michigan 49001) and Robert T. Blackburn.

Changes in Students' Attitudes as a Result of Participating in a Simulated Game. W. Harvey Hegarty (Graduate School of Business, Indiana University, Bloomington, Indiana 47401).

Differences and Changes among College Classes (Freshman through Graduate Levels) in Self-ratings of a Quality Student. Bernadette M. Gadzelle (Department of Psychology, East Texas State University, Commerce, Texas 75428) and Glenn P. Fournet.

(*Continued on p. 973*)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 911-918

INFLUENCE OF CULTURAL BACKGROUND ON HIERARCHICAL LEARNING

RUSSELL D. LINKE[1]

Monash University, Clayton, Australia

This study researched the cross-cultural validity of a learning hierarchy of graphical interpretation skills with 192 Grade 7 students in Australia and 200 indigenous Grade 9 students in Papua New Guinea. Both groups contained approximately equal numbers of male and female students, with a mean age of 12.5 and 15.5 years, respectively for Australia and Papua New Guinea. The learning hierarchy was constructed by logical task analysis, and validated by means of a comprehensive instructional and testing program. Each pair of hierarchically related skills was tested for significance ($p \simeq .001$) and power with a modified version of the R. T. White and R. M. Clark test of inclusion. Results indicate that the pattern of acquisition of interpretative skills was substantially the same for both groups of students, irrespective of their different cultural backgrounds, providing important cross-cultural evidence for the principle of hierarchical learning.

The principle of hierarchical learning involves the sequential acquisition of logically related and progressively complex capabilities, with a fundamental condition that the learning of any particular skill is dependent on the prior learning of specified prerequisite or subordinate skills. This principle is more or less consistent with many recent models of learning and instruction, including those of Gagné (1965), Ausubel (see Ausubel & Robinson 1969 pp. 37, 59), Bloom (1956, p. 30), and Merrill (see Tennyson & Merrill, 1971), but it is probably presented most explicitly in that proposed by Gagné. The logical foundation and structural simplicity of Gagné's model have made it particularly suitable for experimental research, and a number of validation studies, based on the model established by Gagné and Paradise (1961), have been conducted in recent years. These studies have so far been concerned, however, with basic methodological problems involved in the construction and validation of learning hierarchies, as indicated in recent reviews by Walbesser and Eisenberg (1972) and White (1973a).

Many of the more important methodological problems inherent in the earlier studies, recently discussed by White (1973a) and Linke (1973), have now been substantially resolved, and a considerable amount of evidence established in support of Gagné's model. This evidence, however, has largely been derived from isolated validation studies in Australia and, more particularly, in the United States of America, with no previous attempt at replication in other geographical areas, or with students from a different type of cultural environment. Thus, prior to this research, it had yet to be established that Gagné's model of hierarchical learning, or the principle on which this model was based, was in fact a comprehensive one, independent of cultural background. The present research was thus intended to provide a foundation of cross-cultural evidence for Gagné's model, through the construction and validation of a common learning hierarchy of graphical interpretation skills with two groups of students from widely different cultural and educational backgrounds.

[1] The author wishes to thank P. J. Fensham and R. T. White of Monash University, Australia, for their interest and helpful advice throughout the course of this research.

Requests for reprints should be sent to Russell D. Linke, who is now at School of Medicine, Flinders University, Bedford Park, South Australia, 5042 Australia.

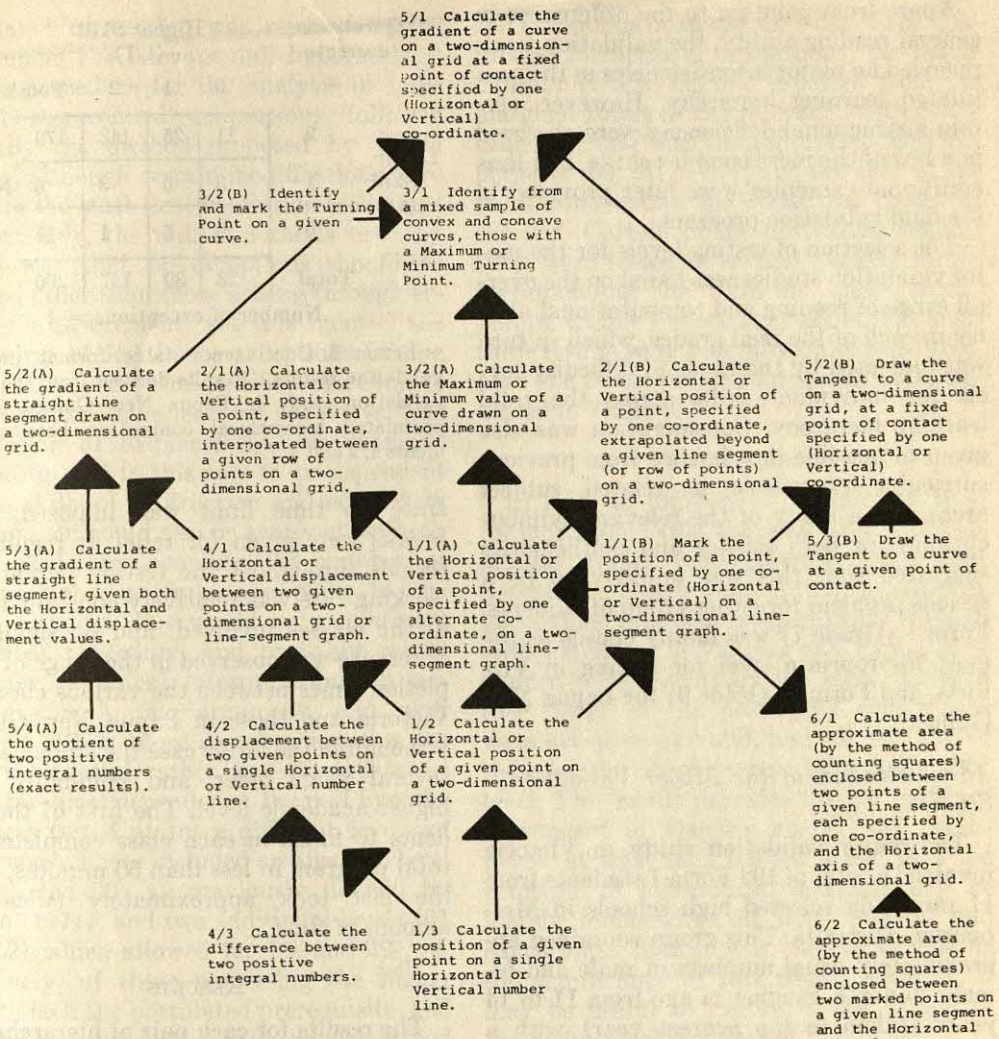
Victoria (Australia) and Papua New Guinea were chosen as suitable areas for the two comparative validation studies. This choice was made partly for pragmatic reasons (Papua New Guinea being at that time an Australian protectorate with a common educational language, and thus relatively accessible for comparative educational research) and partly because of the well-established differences in cultural and educational background between Australian students and those in Papua New Guinea (Prince, 1969; Ralph, 1968). It was therefore expected that a common validation result with students from these two areas would provide a valuable basis of cross-cultural evidence for the principle of hierarchical learning.

Construction and Trial Validation of the Learning Hierarchy

A comprehensive learning hierarchy of graphical interpretation skills (see Figure 1) was constructed by the method of task analysis initially proposed by Gagné and Paradise (1961), then checked for logical consistency by a number of curriculum experts. An extensive analysis was subsequently made of possible subdivisional skills within each of these basic interpretative abilities. For example, the ability to calculate coordinate position on a two-dimensional grid (Element 1/1[A] in the postulated learning hierarchy) was examined for possible differences arising from axis orientation (horizontal or vertical), numerical value (integral or decimal), or sign (positive or negative). This analysis of subdivisional skills was intended to define more precisely the limits of lateral transfer inherent in each of the basic skills, and thus to avoid invalidating likely connections through comparison of different subdivisional skills at successive levels of the learning hierarchy. This procedure was initially used by White (1971), although the method of statistical analysis for differentiating independent subdivisional skills was later modified by Linke (1973).

Following the analysis of subdivisional skills, certain modifications were made to the postulated learning hierarchy, so that each basic skill was defined within a similar range of lateral transfer (subdivisional) conditions. Appropriate instructional and testing materials were subsequently prepared for a trial validation study. These materials were presented in programmatic form, as proposed in an earlier study by White (1971), so that each of the basic skills in turn was defined with appropriate instructions, illustrated with a fully worked example, and finally examined with two analogous questions. The sequence of presentation generally followed a linear pattern, progressing from simpler to more complex skills along each strand of the learning hierarchy, with revisionary questions included at each of the branching points. A limited validation trial was then conducted in Victoria with a single class of students at each of three consecutive academic levels (Grades 6–8). This trial was intended as a practical check on the presentation sequence of skills. It also provided general information on the applicability of instructional materials for each of the three selected grades, from which the most appropriate level for the major validation study could then be determined. A similar trial was conducted with the same validation program in Papua New Guinea, so that any problems arising specifically with students in that area could be examined, and if possible rectified in the major study.

The most obvious result emerging from the validation trials was the difference between students in Victoria and Papua New Guinea with respect to general reading ability. This appeared to be substantially lower for students in Papua New Guinea than for those at equivalent grade levels in Victoria. To some extent this result had been anticipated, since similar differences in reading ability, and in the understanding of common scientific terms, had previously been reported by Gardner for secondary students in Victoria (Gardner, 1972) and Papua New Guinea (Gardner, 1971). These results had been considered in preparing the validation program, so that any terms obviously unfamiliar to the students, such as "gradient," "axis," "tangent," and "displacement," were explained in more familiar terms and illustrated by reference to appropriate examples.



Basic Skills of Graphical Interpretation - An Outline of the Postulated Learning Hierarchy

Classification Code

The first number for each element represents the interpretative area, and the second number indicates the level within that area, rated downward from the relevant terminal skill. Where letters are also used, these indicate secondary sequences within the same interpretative area.

Interpretative Area

1. Position (Location of Co-ordinate Points)
2. Position (Interpolation and Extrapolation)
3. Position (Classification of Turning Points)
4. Displacement
5. Gradient
6. Area

FIGURE 1. Basic skills of graphical interpretation—an outline of the postulated learning hierarchy.

Apart from pointing to the differences in general reading ability, the validation trials revealed no major inconsistencies in the postulated learning hierarchy. However, certain instructional deficiencies were observed in a few of the more complex skills, and thus additional examples were later provided in the final validation program.

The selection of testing levels for the major validation studies was based on the overall range of reading and computational ability in each of the trial grades, which in turn was indicated by the range of difficulty levels for individual skills across the total learning hierarchy. Consideration was also given to possible interference from previous curricular experience in related subject areas, but a study of the relevant syllabus outlines for both science and mathematics suggested that this was unlikely to be a serious problem for either group of students. Form 1 (Grade 7) was finally chosen as the most appropriate level for testing in Victoria, and Form 3 (Grade 9) for Papua New Guinea.

Administration of the Major Validation Studies

The major validation study in Victoria involved a total of 192 Form 1 students from 11 randomly selected high schools in Melbourne, Australia. This group contained approximately equal numbers of male and female students, ranging in age from 11 to 14 years (taken to the nearest year) with a mean of 12.5. The corresponding study in Papua New Guinea involved a total of 200 indigenous Form 3 students from all six central district high schools following the national curriculum. This group, which represented almost half the total number of indigenous students at that level in the central district, contained 139 male and 61 female students, the combined group ranging in age from 13 to 19 years, with a mean of 15.5.

The administrative conditions were similar for both of the major studies. The validation program, which contained both instructional and testing materials, was divided into three convenient sections, each dealing with a different set of skills. These sections were presented consecutively, and since no time limit was imposed, short breaks were given for relief at regular intervals throughout the testing period. The working rate was relatively constant for all of the classes involved, and no significant difference was observed in the range of completion times between the various classes in Victoria and those in Papua New Guinea, although in the latter case the students were several years older, and nominally at a higher academic level. The first of the students to finish in each class completed the total program in less than 60 minutes, while the last took approximately twice this amount of time.

Questions correct		Higher Skill (4/1) 0	1	2	Total	
Lower Skill (4/2)	2	11	25	143	179	Number of students
	1	3	0	3	6	
	0	9	5	1	15	
	Total	23	30	147	200	

Number of exceptions = 1

FIGURE 2. Contingency table format used for validation results. (Results derived from the major validation study in Papua New Guinea for the postulated hierarchical connection between elements 4/2 and 4/7.)

RESULTS

The results for each pair of hierarchically related skills were arranged in a 3 × 3 contingency table (see Figure 2), with the number of questions correct for each individual skill represented by the appropriate marginal totals. Thus the number of exceptions observed for each of the postulated hierarchical connections—that is, the number of students with neither question correct for the lower skill, and both correct for the higher skill—was indicated in the bottom, right-hand (0/2) cell. This number was then compared with an expected maximum or critical number, calculated from the various marginal totals according to the method proposed by White and Clark (1973), and if the actual number of exceptions observed was greater than that excepted, then the

postulated connection was rejected as invalid under the relevant null hypothesis.

This procedure for the analysis of individual hierarchical connections followed basically the method proposed by White (1971), although certain modifications were made to the statistical validation technique. For example, the null hypothesis used by White was that no exceptions should be allowed other than those arising through errors of measurement, yet this ignored the possibility of legitimate exceptions arising through the use of unidentified prerequisite skills or alternative learning pathways. The calculation of displacement, for example, was postulated in this case to be dependent on the skills of subtraction and coordinate location, but could also be achieved, without these skills, by counting the number of marked divisions from one point to the other, although this would be a relatively inefficient procedure, and probably rarely used for graphical interpretation. In accordance with this argument, and with Gagné's (1970) more recent orientation toward substantial, rather than absolute levels of hierarchical dependence, the null hypothesis was tested at three different levels of stringency. These included an absolute level (designated 00) as previously defined by White (1971), and two additional levels (01 and 02) which allowed for 1% and 2%, respectively, of those possessing the higher skill to lack the postulated prerequisite, and thus to have used alternative or unspecified subordinate skills. The power in each case was calculated against the same alternative hypothesis, which in turn was defined to allow a much higher level of exceptions (10%), thus indicating no substantial relationship of hierarchical dependence.

The overall level of significance was arbitrarily set at .05, so that the probability of a Type I error (false rejection of the null hypothesis) for each of the 61 tests (including parallel connections for different subdivisional conditions) in each of the major validation studies was determined by $\alpha = 1 - \sqrt[61]{1 - .05} \simeq .001$ (Linke, 1973).

A summary of results for both of the major validation studies is presented in Table 1. These results indicate the number of 0/2 cell exceptions observed for each of the postulated hierarchical connections, the maximum or critical number expected from the marginal totals of the relevant contingency table, together with the appropriate null hypothesis (00, 01, or 02), and for each of the valid connections the statistical power of the test. Certain of the postulated connections were tested in several parallel forms, corresponding to different subdivisional groups within the relevant skills, but since these results were generally quite consistent, only one of the parallel forms has been presented in Table 1.

Discussion

In spite of differences in difficulty level for individual skills, the results for both validation studies were generally quite consistent, and clearly showed that the learning hierarchies validated independently for high school students in Victoria and Papua New Guinea were essentially the same. In each case, most of the postulated connections were accepted as valid, and all but one of these at the absolute null hypothesis (00) level. This result provides further evidence in support of Gagné's model, and establishes, at least in a limited sense, that the principle of hierarchical learning may be largely independent of both cultural and educational background. Before discussing the significance of this result, however, it may be useful to explain a few points of methodological importance.

The occasional inconsistencies observed in these results, for example the postulated connections between skills 1/2 and 1/1(B), 1/1(A) and 2/1(B), and 1/1(A) and 3/2(A), which were each rejected at the weakest null hypothesis level in one of the validation studies, occurred through incidental acquisition of two positional skills (1/2 and 1/1(A)) in the course of subsequent attempts at more complex capabilities. An analysis of test–retest results for the two anomalous skills confirmed this hypothesis of incidental acquisition, and indicated that the problem was caused by repetition of certain instructional procedures for each of the relevant superordinate skills, which effectively provided an additional op-

TABLE 1
OUTLINE OF VALIDATION RESULTS

Lower skill	Higher skill	Victoria			Papua New Guinea		
		Number of Exceptions		Power	Number of Exceptions		Power
		Observed	Expected[a]		Observed	Expected[a]	
1/3	1/2	4	6/01	.97	3	3/00	1.00
1/2	1/1(B)	15	12/02	—	5	7/00	.97
1/1(B)	1/1(A)	0	7/00	.82	0	10/00	.84
1/2	1/1(A)	9	9/02	.49	4	6/00	.97
1/1(A)	2/1(A)	14	12/02	—	19	14/02	—
1/1(A)	2/1(B)	8	8/00	.41	17	11/02	—
1/1(A)	3/2(A)	13	12/02	—	7	9/00	.73
3/2(A)	3/1	4	15/00	.02	11	20/00	.01
3/2(B)	3/1	2	7/00	.12	1	9/00	.01
4/3	4/2	0	2/00	1.00	0	1/00	1.00
1/3	4/2	0	1/00	1.00	1	2/00	1.00
4/2	4/1	1	5/00	.97	1	9/00	.90
1/2	4/1	0	6/00	.95	2	8/00	.93
5/4(A)	5/3(A)	2	3/00	1.00	1	4/00	1.00
5/3(A)	5/2(A)	3	3/00	.58	0	2/00	.93
4/1	5/2(A)	1	3/00	.59	3	3/00	.86
5/3(B)	5/2(B)	0	6/00	.74	1	8/00	.87
1/1(B)	5/2(B)	2	4/00	.90	0	3/00	1.00
5/2(B)	5/1	0	6/00	.01	0	9/00	.01
5/2(A)	5/1	0	7/00	.01	1	11/00	.01
6/5(B)	6/4(B)	0	2/00	1.00	0	2/00	1.00
6/4(B)	6/3(B)	4	4/00	.98	1	2/00	1.00
6/3(B)	6/2	0	5/00	.43	0	8/00	.36
6/3(A)	6/2	0	7/00	.29	0	11/00	.18
6/2	6/1	2	6/00	.15	2	8/00	.04
1/1(B)	6/1	0	2/00	.64	0	1/00	.50

[a] The second number in this column indicates the appropriate null hypothesis.

portunity to learn the postulated prerequisites. Whether these results should be classified as legitimate exceptions to the postulated learning hierarchy is an arguable point. It is possible, for example, as suggested by White (1971), that the subordinate skills might in fact have been acquired first within the context of the higher skill, and then immediately used to solve the more complex questions. Thus, any students succeeding at the higher skill, and subsequently found to have acquired incidentally the postulated subordinate skills, could be eliminated from the relevant validation test for want of definitive information on which of the skills was actually first acquired (White 1971). It could also be argued, however, that any skill acquired incidentally should not be classified as an independent prerequisite ability, and this leads in turn to an unresolved argument of basic definition. It suffices to conclude for this study, however, that the validation test was conservative in accounting for all exceptions to the postulated learning hierarchy.

Another source of inconsistent results, and in fact one of the most significant problems with this validation technique, was the method of calculating statistical power. In some of the postulated connections, for example, the total number of students who failed the subordinate skill was lower than the calculated critical number of 0/2 cell exceptions, so that the relevant connection could not possibly be rejected. In cases such as this, the power, or probability of detecting a genuinely invalid connection, should have been extremely low, but since the alternative hypothesis took no account of subordinate-skill difficulty levels, the resultant levels of power were unrealistically high. A similar problem occurred in the previous

validation study by White (1971), and was caused by the same inadequate formulation of the alternative hypothesis. In contrast with White's formulation, however, the alternative hypothesis was related in this case to the number of students who succeeded at the superordinate skill, so that high difficulty levels, as expected, reduced the effective difference between null and alternative hypotheses, and hence also drastically reduced the statistical power. Unfortunately, this situation was often complicated by large numbers of experimental errors, both for guessing and chance mistakes, occurring in many of the more difficult skills. Although this increased the critical number of 0/2 cell exceptions, and hence the probability of accepting the relevant connection, it also reduced substantially the power of the test. There is some consolation in these results, however, that the observed number of exceptions was generally much lower than that expected, and that the results for both validation studies, and indeed all parallel tests for different subdivisional skills, were invariably consistent.

Methodological issues aside, the predominantly similar patterns of hierarchical learning, independently established in parallel studies for high school students in Victoria and Papua New Guinea, have provided an important foundation of cross-cultural evidence for Gagné's model of hierarchical learning. The significance of this result is probably enhanced by the fact that it covers a difference of several years in age and nominal academic level, and substantial differences in cultural and educational background. Moreover, the difference in cultural background between Australian students and those in Papua New Guinea appears to be particularly pronounced for quantitative intellectual skills of the type defined in this research. Prince (1969) observed, for example, that Western scientific concepts were "almost nonexistent" in the indigenous cultures of Papua New Guinea, and that mathematical concepts were "even more rudimentary." Johnson (1968) also observed that concepts such as quantity, space, and time were "not expressed in indigenous New Guinean culture or its languages [p. 21]," and King (1970) has more recently suggested that children in Papua New Guinea might not form mathematical concepts "either in the same order or way, or at the same chronological age as their counterparts in modern Western society [p. 58]." This research has established at least that the pattern of acquisition for certain basic skills of graphical interpretation is substantially the same.

The difference in formal educational background between Australian students and those in Papua New Guinea is much less pronounced than that of cultural background, but nevertheless important. Although English, for example, is the common educational language in both areas, for indigenous students in Papua New Guinea it is very much a second language, rarely spoken outside the classroom. In addition, both primary and secondary curriculum studies in Papua New Guinea, although essentially prepared by Australian and overseas expatriate teachers, have been devised with some consideration for the characteristic needs of a relatively primitive and segregated rural society (Ralph, 1968). Thus, very few indigenous students, and none of those involved in this research, would have experienced a similar course of study to that used in Victorian schools.

In spite of these differences in curricular background, however, it must be admitted that more than half of the secondary teachers in Papua New Guinea are Australians or overseas expatriates, as are many of the primary teachers and most of the teacher training staff (Education Department of Papua New Guinea, 1973). Moreover, the use of stringent academic selection procedures, which in turn cause a drastic reduction in the number of students continuing from primary to lower secondary levels, means that the Form 3 students involved in this study probably represent, in the Western sense, an academic elite among the indigenous population. It could therefore be argued that the common pattern of learning observed in this research might simply be a product of successful acculturation for the indigenous students of Papua New Guinea in a basically Western system of formal ed-

ucation. There is no complete counter argument, since it would obviously be impossible to conduct a strictly comparative validation study in any of the many vernacular modes of Papua New Guinea. However, the evidence provided by Prince (1967) and Mackay and Gardner (1969), concerning characteristic difficulties with scientific concepts and quantitative skills among relatively senior and academically successful students in Papua New Guinea, emphasizes the overriding influence of traditional culture on the learning of complex and formalized intellectual skills, and therefore counteracts, at least to some extent, the argument of successful acculturation.

It seems appropriate at this stage to make two cautionary notes regarding the potential generalizations of this result. In the first place, learning hierarchies of the type defined in this research may only be appropriate for generalized or intellectual skills (White, 1973b), and other forms of learning, such as memorization (Tennyson & Merrill, 1971), rote learning (Ausubel & Robinson, 1969), or simple recall of verbal information, may not conform to the same basic principle of hierarchical acquisition. Second, the validation process used in this research was intended only to examine immediate application of the relevant intellectual skills, and did not incorporate any test of long-term meaningful learning. In spite of these potential limitations, however, the cross-cultural validation of a common learning hierarchy of graphical interpretation skills for students in Australia and Papua New Guinea has made an important contribution of empirical evidence to support the basic principle of hierarchical learning.

REFERENCES

Ausubel, D. P., & Robinson, F. G. *School learning: An introduction to educational psychology*. New York: Holt, Rinehart & Winston, 1969.

Bloom, B. S. (Ed.) *Taxonomy of educational objectives—The classification of educational goals. Handbook 1: Cognitive Domain*. London: Longmans, 1956.

Education Department of Papua New Guinea. *Official statistics*. Territory of Papua New Guinea: Author, 1973.

Gagné, R. M. *The conditions of learning*. New York: Holt, Rinehart & Winston, 1965.

Gagné, R. M. Some new views of learning and instruction. *Phi Delta Kappan*, 1970, **51**, 468–472.

Gagné, R. M., & Paradise, N. E. Abilities and learning sets in knowledge acquisition. *Psychological Monographs*, 1961, **75** (14, Whole No. 518).

Gardner, P. L. Project SWNG: Scientific words—New Guinea. (An investigation of non-technical vocabulary difficulties amongst Form III and IV science students in the Territory of Papua and New Guinea). (Tech. Rep.) Territory of Papua-New Guinea: Department of Education, 1971.

Gardner, P. L. Words in science: An investigation of non-technical vocabulary difficulties amongst Form I, II, III, and IV science students in Victoria. (Special Report) Melbourne, Victoria: Australian Science Education Project, 1972.

Johnson, F. C. The role of English in Papua and New Guinea. *Papua and New Guinea Journal of Education*, 1968, **5**, 16–26.

King, B. W. Objectives in mathematics: Why teach set theory? *Papua and New Guinea Journal of Education*, 1970, **6**, 54–66.

Linke, R. D. The effects of certain personal and situational variables on the acquisition sequence of graphical interpretation skills. Unpublished doctoral dissertation, Monash University, Victoria, 1973.

Mackay, L. D., & Gardner, P. L. Conceptual and vocabulary difficulties in physical science amongst Papuan and New Guinea high school students. *Papua and New Guinea Journal of Education*, 1969, **6**, 3–11.

Prince, J. R. Bottlenecks in Papua New Guinea Science Teaching. *Papua and New Guinea Journal of Education*, 1967, **5**, 17–26.

Prince, J. R. *Science concepts in a Pacific culture*. Sydney: Angus & Robertson, 1969.

Ralph, R. C. Education in Papua-New Guinea: Integration—whither? *Australian Journal of Education*, 1968, **12**, 30–45.

Tennyson, R. D., & Merrill, M. D. Hierarchical models in the development of a theory of instruction. *Educational Technology*, 1971, **11**, 27–31.

Walbesser, H. H., & Eisenberg, T. A. A Review of Research on Behavioural Objectives and Learning Hierarchies. (Mathematics Education Rep.) Columbus, Ohio: ERIC Information Analysis Center for Science, Mathematics and Environmental Education, 1972.

White, R. T. Learning graphical skills in kinematics. Unpublished doctoral dissertation, Monash University, Victoria, 1971.

White, R. T. A limit to the application of learning hierarchies. *Australian Journal of Education*, 1973, **17**, 153–156. (a)

White, R. T. Research into learning hierarchies. *Review of Educational Research*, 1973, **43**, 361–375. (b)

White, R. T., & Clark, R. M. A test of inclusion which allows for errors of measurement. *Psychometrika*, 1973, **38**, 77–86.

(Received February 18, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 919-925

OUTERDIRECTEDNESS IN CHILDREN OF THREE AGES AS A FUNCTION OF EXPERIMENTALLY INDUCED SUCCESS AND FAILURE

DONALD L. MacMILLAN[1] AND DEBORAH L. WRIGHT[2]

University of California, Riverside

Studied the problem-solving style of outerdirectedness in children from three grades under two conditions—experimentally induced success and failure. The total sample of 60 subjects were stratified by grade (second, fourth, and sixth) and sex. Subjects in each of the groups thus created were randomly assigned to treatment conditions. Success and failure were experimentally induced by means of a puzzle task, and subjects were then required to perform two dependent tasks (a puzzle task and a sticker game imitation) from which outerdirectedness was inferred. Results supported the hypothesis that children shift from an outerdirected to an innerdirected problem-solving style as age increases. No sex differences were found on any dependent measure. The hypothesized effect of failure resulting in greater outerdirectedness received partial support, in that subjects evidenced more outerdirected behavior under failure conditions on the criterion puzzle task. This was possibly due to the similarity between experimental and criterion tasks. Several methodological issues pertaining to experimental induction of failure, the task-specific nature of outerdirectedness, and the validity of the dependent tasks for assessing outerdirectedness were discussed.

Outerdirectedness is defined as a style of problem solving characterized by a reliance on concrete situational cues rather than by active attempts to deduce abstract relationships. In essence, this means that the individual whose problem solving is characterized as outerdirected relies unduly on external cues resulting in little reliance on his own cognitive resources.

The vast majority of research on outerdirectedness has focused upon mentally retarded subjects (see Sanders, Zigler, & Butterfield, 1968; Turnure, 1970a, 1970b, 1970c, 1973; Turnure & Zigler, 1964), since outerdirectedness as a construct first arose out of a series of studies by Zigler (1966) and his associates which grew out of motivational explanations for what previously had been considered rigidity (Kounin, 1962) in the retarded. Outerdirectedness is thought to be related to two factors: the level of cognition attained, or mental age (MA), and the degree of success experienced through employing whatever cognitive resources a child has available (Turnure & Zigler, 1964). That is, the lower the MA, the more outerdirected one would be, and the more failure an individual has encountered in problem solving, the more he comes to distrust his own solutions and search for cues in the environment as guides. Hence, outerdirectedness is not an inherent part of retardation but rather it is seen as an outgrowth of the excessive failure encountered by the retarded. The evidence to date on the role of MA (Achenbach & Zigler, 1968; Massari & Mansfield, 1973; Yando & Zigler, 1971) is inconclusive with only Yando and Zigler (1971) demonstrating an MA effect. The role of failure has been well documented (Green & Zigler, 1962; Turnure, 1970a, 1970b, 1970c; Turnure & Zigler, 1964).

Children with histories of success (such as the children serving as subjects in the

[1] Requests for reprints should be sent to Donald L. MacMillan, Department of Education, University of California, Riverside, California 92502.

[2] Now employed by the Alvord Unified School District, Riverside, California. This research was supported in part by U. S. Office of Education Grant OEG-0-73-5263. The authors wish to thank A. B. Silverstein for his assistance in the data analysis.

present study) are thought to shift developmentally from outerdirected to innerdirected problem-solving strategies (Achenbach & Zigler, 1968: Turnure, 1973). One question tested by the present study was whether exposure to experimentally induced failure would result in normal children reverting back to an outerdirected style as they perceive their own attempted solutions to the tasks as inadequate.

In the research conducted to date, nonorienting behavior is descriptive of when the subject looks away from the experimental task. For example, in the Turnure and Zigler (1964) study, the subject was to assemble one puzzle while the experimenter assembled a second puzzle. Observation of the experimenter's activity could be helpful, detrimental, or of no consequence to the subject on a subsequent task to be assigned. The inference that the nonorienting behavior reflected "outerdirectedness" rather than "distractibility" depends on whether the second experimental task is performed more rapidly (or more slowly if the experimenter activity was irrelevant to the solution of the second task) than the first task and/or more rapidly than a control group in which no experimenter activity was evidenced.

In a recent study (MacMillan & Cauffiel, 1973), the outerdirected construct was extended to children of average ability with academic histories of failure (i.e., variously called underachievers, or children with learning disabilities). The results of that study were supportive of the role of failure in determining outerdirectedness but raised certain questions regarding the task-specific nature of outerdirectedness. That is, whether a subject is deemed outerdirected in his problem solving may depend on the specific nature of the dependent task used to measure outerdirectedness. These methodological problems are considered in the present study.

Method

Subjects

A total sample of 60 subjects was stratified by grade level. Samples of 10 males and 10 females were drawn at random from two schools in a single urban school district from among each of the following grade levels: second, fourth, and sixth (chronological age, CA, ranges of 7 years 11 months to 8 years 10 months, 9 years 11 months to 10 years 10 months, and 11 years 11 months to 12 years 10 months, respectively). The school district in which the study was conducted is one of three districts in an urban area and serves primarily lower-middle-class families. No children were used who displayed any motor, perceptual, or sensory impairments, nor were any children used who received any specialized remedial help. Three male and three female subjects had Spanish surnames with the remainder of the subjects being Anglo. Table 1 contains means and standard deviations for CA by groups. Hence a 3 × 2 (Grade Level × Sex) sampling design was created.

TABLE 1
Chronological Age of Subjects

Sex	Grade level		
	2nd	4th	6th
Male			
M	99.00	125.00	150.20
SD	1.76	3.17	1.87
Female			
M	99.20	128.40	149.80
SD	2.04	3.52	1.69

Note. $N = 60$.

Procedures

Children were seen individually in their own school, on one occasion by a female examiner who was unaware of the hypotheses being tested. Subjects within each cell (e.g., second-grade males) were assigned at random to one of two treatments: success or failure.

Interpolated success. Three pictures (a duck, a sailboat, and an airplane) were divided into three pieces. In a pilot study, these designs had been found to be easily recognized as to what they were a part of, and they were easy to assemble for children in the age range of subjects used in this study. Under the success condition, the experimenter allowed the subject to complete the assembly of the puzzle and then clicked the stopwatch. The experimenter then removed the completed puzzle and said, "That was very good. You're good at putting puzzles together." Then the second puzzle was given to the subject and again he was allowed to complete it, whereupon the examiner said, "You did very well on this puzzle, could you put another one together for me?" The third puzzle was then presented and upon completion, the child was told, "You put this puzzle together better than anyone I have asked. You're very good at this."

Interpolated failure. Those subjects selected to receive the failure treatment were not allowed to complete any of the three puzzles. The puzzles used with the failure subjects were divided into

20 pieces each, and the child was given one minute in which to assemble it. Data from a pilot study revealed that these puzzles could not be assembled in one minute by children of the ages used in this study. After the subject was stopped on the first puzzle, the examiner said, "You did not finish the puzzle. You should have been able to finish it before your time was up. Since you didn't finish I'll give you another puzzle." Then the second puzzle was presented, and when the child was stopped again, he was told, "Well, I see you did not finish it before the time was up. You didn't finish this one either." Finally, the third puzzle was attempted and again the child was stopped prior to completing the puzzle, whereupon he was told, "You did not do very well on these puzzles. You must not be very good at putting puzzles together. All the other children I asked to put these together did them correctly before the time was up."

Dependent measures. Following the administration of success or failure protocols, each subject was then given two tasks: puzzle task and sticker game task (after Turnure & Zigler, 1964). The order of these tasks was randomly determined for each subject.

The puzzle task consisted of two puzzles (adapted from the horse and elephant from the Wechsler Intelligence Scale for Children, WISC, Object Assembly). The experimenter explained, "Here are some pieces of a puzzle. When you put them together, they will make something you know. I want you to put them together as quickly as you can. While you are putting yours together, I will put one together too. But you put yours together as fast as you can. Any questions? Okay, here's your puzzle. Begin."

As soon as the subject began working, the experimenter started the stopwatch. While the subject assembled the first puzzle (which was randomly determined by the subject), the experimenter assembled the other puzzle. The experimenter left the completed puzzle in view of the subject for 10 seconds, then disassembled it and left the pieces in view for 30 seconds. If the subject had not completed work on the first puzzle, the experimenter repeated the cycle with the puzzle.

Two dependent measures were recorded while the subject worked on the first puzzle task: (*a*) time required to assemble the puzzle correctly and (*b*) the number of times the subject glanced at the experimenter or the experimenter's puzzle.

The subject was then given the second puzzle (the one the experimenter assembled while the subject assembled the first puzzle) and told him, "Here is another puzzle to put together as quickly as you can. Do it as fast as you can. Any questions?" While the subject assembled the second puzzle, time and number of glances were again recorded.

A second set of dependent measures were derived from the sticker game task, which followed the protocols of Turnure and Zigler (1964). In this task, the examiner opened a box of stickers of several colors (black, green, red, and yellow) and took a sheet of white paper. The experimenter then explained, "I'll make a design out of these pieces of sticker paper on this white sheet first, then you may make a design." The experimenter then made one of three predetermined designs with one of the colors of stickers (again randomly determined). Upon completion of the design, the experimenter named her design, "I'll call my design a __ __ (either a bow-wow dog, a pretty tree or a go-go-cart). Now you may make any design you wish, with any color paper. After you have finished, you may give your design a name which I will write at the bottom of your design." This procedure was repeated for the other two designs.

Hence, from the sticker game task the following measures were taken on each of the three designs: (*a*) the color of stickers used, (*b*) the design itself, and (*c*) the name given the design by the subject. One point was given for each of these three elements if they corresponded to the color, design, and name used by the experimenter, making a total of nine points possible over the three designs.

Following the completion of the testing, all subjects were debriefed and told that the interpolated success and failure were not a true reflection of their performance on the preexperimental puzzles.

Results

Each of the dependent measures was analyzed separately in order to evaluate differences due to grade levels, treatments, sex, and the interactions thereof. Data for the first two measures, sticker game and glancing, were subjected to a three-factor analysis of variance (Grade × Treatment × Sex). The third measure, time spent on the puzzles, was subjected to a four-factor analysis of variance with repeated measures on the fourth factor, puzzles. The .05 level of significance was adopted for all statistical tests. Summary data for all dependent measures are shown in Table 2.

The hypothesized difference for grade level was supported by all three dependent measures. Significant differences were found for all dependent measures (sticker game, $F = 4.81$, $df = 2$, $p < .05$; glancing, $F = 3.96$, $df = 2$, $p < .05$; time on puzzles, $F = 6.24$, $df = 2$, $p < .05$). Post hoc tests were performed according to Tukey's test for honestly significant differences (HSD) and revealed a significant difference between second-grade and sixth-grade subjects on both the sticker task ($q = 4.36$, $p < .05$) and number of glances ($q = 3.92$, $p < .05$). On the time required to complete the puzzles, it was found that sixth-grade sub-

TABLE 2

MEANS AND STANDARD DEVIATIONS FOR ALL DEPENDENT MEASURES BY GRADE, SEX, AND TREATMENTS

Grade/sex	Treatment	Time						Glancing		Sticker game	
		Puzzle 1		Puzzle 2		Total					
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
2nd											
Male	success	27.20	11.69	17.00	6.60	44.20	16.24	1.20	1.30	2.40	1.14
	failure	60.40	37.78	63.20	60.87	125.60	62.42	2.60	1.52	2.60	1.52
Female	success	65.00	57.92	27.20	12.87	92.20	59.98	2.20	1.79	3.00	1.41
	failure	53.60	32.05	31.60	32.83	85.20	41.09	3.00	2.55	3.20	.84
4th											
Male	success	20.00	9.35	14.00	4.42	34.00	11.38	.20	.45	1.60	1.34
	failure	68.40	61.65	62.20	83.66	130.60	96.78	1.40	2.07	2.00	1.00
Female	success	68.00	51.43	47.60	41.79	115.60	75.99	1.80	1.79	1.80	.84
	failure	44.40	59.37	46.00	45.74	90.40	65.49	1.80	2.95	3.00	1.73
6th											
Male	success	22.60	20.01	9.60	2.40	32.20	20.98	.40	.55	1.40	.89
	failure	37.20	30.61	11.00	4.53	48.20	35.11	1.60	.89	1.40	.89
Female	success	25.40	17.29	13.80	6.30	39.20	22.06	.20	.45	2.00	.71
	failure	27.40	15.66	14.40	6.80	41.80	21.00	1.00	1.22	1.60	1.82

jects took significantly less time to complete the puzzles than both second-grade and fourth-grade subjects. In short, the developmental hypothesis, that as children get older they become less reliant on external cues, was supported by the results for all three dependent measures.

The hypothesized effect of interpolated failure resulting in more nontask orientation received partial support. On the sticker task, the effect of treatments failed to reach statistical significance. However, on both dependent measures derived from the puzzle task (glances and time), a significant effect for treatments was found (glancing, $F = 4.43$, $df = 1$, $p < .05$; time, $F = 4.21$, $df = 1$, $p < .05$). In both instances, interpolated failure resulted in subjects glancing more frequently and taking more time to complete the puzzles than was the case for subjects performing under the interpolated success condition.

The third hypothesis, that girls would be more outerdirected than boys, failed to receive support from any of the dependent measures. However, on one dependent measure (time to complete tasks), a significant interaction between treatment and sex ($F = 7.84$, $df = 1$, $p < .05$) was found. Means and standard deviations by cells are shown in Table 3 for this interaction. Post hoc tests revealed that males took significantly longer to complete puzzles under the failure condition than under success conditions ($t = 3.43$, $p < .05$), and females took significantly longer than males under the success condition ($t = 2.44$, $p < .05$).

In order to test for a relationship among dependent measures, a chi-square test was run. The decision to use a nonparametric test was based on the fact that the scores derived from both glancing and the sticker game were ordinal at best and the distributions were not continuous. Since all three dependent measures used in this study have been used either singly or in combination in previous studies, it was judged important to

TABLE 3

MEANS AND STANDARD DEVIATIONS BY CELLS FOR TREATMENT × SEX INTERACTION ON TIME TO COMPLETE PUZZLES

Sex	Treatment	
	Success	Failure
Males		
M	36.80	100.80
SD	16.37	74.48
Females		
M	78.66	65.80
SD	64.51	47.52

TABLE 4

CELL FREQUENCIES FOR THE RELATIONSHIP BETWEEN SCORES ON THE THREE DEPENDENT MEASURES

Measure	Glances on Puzzle 1		Sticker game score	
	0–1	2–5	0–2	3 or more
Time on Puzzle 2				
0–13 seconds	19	3	17	5
14–24 seconds	14	7	11	10
25+ seconds	8	9	7	10
Glancing behavior				
0–1			26	15
2–5			9	10

ascertain whether they were tapping related processes. In Table 4 are the cell frequency counts upon which the chi-square was computed on time required to complete Puzzle 2 and glancing behavior, time required to complete Puzzle 2 and sticker game scores, and glancing behavior and sticker game scores. It should be noted that time on Puzzle 2 was used as that score is the one presumably affected by the number of glances (i.e., more glancing at the second task while assembling the first puzzle should increase the speed with which the second puzzle is assembled) during the assembly of the first puzzle. Results revealed a reliable relationship between the time required to complete the second puzzle and glancing behavior ($\chi^2 = 6.87$, $df = 2$, $p < .05$); however, visual inspection of the frequencies in Table 4 reveal that the relationship which was found was the opposite of the one predicted. Namely, subjects who rapidly completed Puzzle 2 tended to be those who glanced very infrequently while assembling the first puzzle. The other analyses failed to uncover a statistically significant relationship between the other dependent measures (time on second puzzle vs. sticker game, $\chi^2 = 5.65$, $df = 2$, $.10 < p > .05$; sticker game vs. glancing, $\chi^2 = .79$, $df = 1$, *ns*). The relationship between time on the second puzzle and sticker game score, while not reaching the adopted level of significance, was again not in the predicted direction. Subjects who were most imitative on the sticker game (i.e., were outerdirected) were not inclined to complete the second puzzle very rapidly.

A follow-up questionnaire provided some additional data of interest. When asked if the puzzles were easy or hard, the vast majority of subjects who performed under the success condition reported that they were easy and indicated they would like to "play the game again." However, those who performed under the failure condition attributed their failure to complete the preexperimental puzzles to either characteristics of the puzzles (e.g., they had too many pieces) or their own inadequacies (e.g., I'm not too good at puzzles). Only 3 of the 15 males under failure conditions blamed the puzzles for the incompletion (all 3 were second-grade children). The responses of females were approximately evenly divided at all three grade levels.

DISCUSSION

The present study extended the outerdirected construct to normal-achieving children. In so doing, an attempt was made primarily to test several assumptions underlying the previous work with mentally retarded subjects. Specifically, the notion that outerdirectedness is a function of failure was tested by means of interpolating success and failure prior to performance on criterion tasks.

In general, the findings of the present study were supportive of the developmental trend reported elsewhere (MacMillan & Cauffiel, 1973; Yando & Zigler, 1971; Zigler & Yando, 1972); namely, that as children get older they become less outerdirected. This finding seems to be a reliable one as Yando and Zigler (1971) used different populations (institutionalized and noninstitutionalized retardates classified as familial and organic) as did Zigler and Yando (1972) whose subjects were institutionalized and noninstitutionalized retarded and normal. The same trend was reported by MacMillan and Cauffiel (1973) with educationally handicapped and normal children. In the present study, the decrease in outerdirected behavior was evidenced on both the sticker game task and the glanding measures, whereas time spent on puzzles

failed to reveal this gradual progression toward less outerdirectedness. In light of the similar findings by MacMillan and Cauffiel (1973), Yando and Zigler (1971), and Zigler and Yando (1972) with the sticker game task, it would appear that this task is sufficiently difficult and/or ambiguous to tap outerdirectedness, particularly since the cognitive demands of this task are minimal.

The findings of the present study, however, raise some doubt as to the validity of the interpretation of greater speed on the second puzzle as indicative of a shift from outerdirectedness to innerdirectedness. First, the nature of the relationship between glancing behavior and time required to complete the second puzzle fails to confirm the notion that the glancing behavior is beneficial. One might also argue that the increased speed of puzzle assembly is simply a function of increased "power" as MA and/or CA increased.

The hypothesized sex difference was not supported by the results of this study. No main effect for sex was found on any of the three dependent measures. One finding of interest, despite the fact that it had not been predicted, was the interaction of Treatment × Sex on the puzzle-time measure. On this task, girls were more outerdirected under success conditions than were boys, and conversely, boys were significantly more outerdirected under the failure condition. This greater outerdirectedness by boys in the failure condition is consistent with previous findings (Achenbach & Zigler, 1968; Turnure, 1970).

Findings related to success and failure treatments were the most intriguing. First, the hypothesized effect of failure increasing outerdirectedness was supported by the results for glancing and puzzle time but not for the sticker game task. It must be noted that the preexperimental tasks with which failure was induced were puzzles, and the two dependent measures for which failure altered performance were both derived from a similar puzzle task (i.e., glancing and time). On the sticker game task, the effect of failure failed to reach significant proportions. This finding underscores the point made by Butterfield and Zigler (1965) that different degrees of similarity between tasks on which success and failure and criterion tasks may account for inconsistent findings between studies which experimentally induce success and failure. A similar transference was noted by MacMillan and Cauffiel (1973) and by Turnure and Zigler (1964). In light of the inconclusive findings with regard to the effect of experimentally induced failure on performance (MacMillan & Cauffiel, 1973; Steigman & Stevenson, 1960; Stevenson & Pirojnikoff, 1958), greater attention need be given to differing methods of inducing failure, transference effects to various dependent tasks, differential perceptions of the reason for failure by populations differing in their histories of failure (MacMillan & Keogh, 1971), and the possibility that a given failure experience may have different potency for populations differing in their histories of failure (MacMillan & Cauffiel, 1973).

The failure to find the predicted relationships between dependent measures raises doubts concerning whether they are tapping related processes. The statistically significant relationship found between glances on Puzzle 1 and the time required to complete Puzzle 2 is of even more concern as it appears to indicate that subjects who glance frequently during Puzzle 1 are not glancing at Puzzle 2 but at stimuli that are either of no value for the immediate task or interfere with the assembly of Puzzle 2. In fact, the present results seem to be more reasonably interpreted as supportive of distractibility than of outerdirectedness. The very tentative finding on the relationship between time on Puzzle 2 and sticker game score seems to indicate that less imitative subjects were faster on Puzzle 2 and the more imitative were slower. Coupling these findings with the failure to find any significant relationships between these same dependent measures with underachieving males (MacMillan & Cauffiel, 1973) questions the validity of these tasks for assessing outerdirectedness.

The follow-up interview used in this study revealed that a majority of boys who were not allowed to complete the puzzles in the preexperimental phase of the study (performing under the failure condition) attrib-

uted their failure to complete the puzzle to characteristics inherent in the puzzle rather than their own inabilities. The three boys who blamed themselves for failing to complete the puzzles were second graders, which is the opposite of the predicted direction of the relationship between internalizing blame and outerdirectedness.

In summary, the findings of the present study highlight the need for further research on three dimensions: the task specific nature of outerdirectedness, the effects of failure as a result of the precise manner in which failure is induced, and the validity of the dependent measures used to assess outerdirectedness.

REFERENCES

Achenbach, R., & Zigler, E. Cue learning and problem-learning strategies in normal and retarded children. *Child Development*, 1968, **39**, 821–847.

Butterfield, E. C., & Zigler, E. The effects of success and failure on the discrimination learning of normal and retarded children. *Journal of Abnormal Psychology*, 1965, **70**, 25–31.

Green, C., & Zigler, E. Social deprivation and the performance of retarded and normal children on a satiation type task. *Child Development*. 1962, **33**, 499–508.

Kounin, J. S. Experimental studies of rigidity: I. The measurement of rigidity in normal and feebleminded persons. In E. P. Trapp & P. Himelstein (Eds.), *Readings on the exceptional child*. New York: Appleton-Century-Crofts, 1962.

MacMillan, D. L., & Cauffiel, S. R. Outerdirectedness as a function of interpolated success and failure in educationally handicapped and regular class males. Unpublished study, University of California, Riverside, 1973.

MacMillan, D. L., & Keogh, B. K. Normal and retarded children's expectancy for failure. *Developmental Psychology*, 1971, **4**, 343–348.

Massari, D. J., & Mansfield, R. S. Field dependence and outerdirectedness in the problem solving of retardates and normal children. *Child Development*, 1973, **44**, 346–350.

Sanders, B., Zigler, E., & Butterfield, E. C. Outerdirectedness in the discrimination learning of normal and mentally retarded children. *Journal of Abnormal Psychology*, 1968, **73**, 368–375.

Steigman, M. J , & Stevenson, H. W. The effects of pretraining reinforcement schedules on children's learning. *Child Development*, 1960, **31**, 53–58.

Stevenson, H. W., & Pirojnikoff, L. A. Discrimination learning as a function of pretraining reinforcement schedules. *Journal of Experimental Psychology*, 1958, **56**, 41–44.

Turnure, J. E. Children's reactions to distractors in a learning situation. *Developmental Psychology*, 1970, **2**, 115–122. (a)

Turnure, J. E. Distractability in the mentally retarded: Negative evidence for an orienting inadequacy. *Exceptional Children*, 1970, **37**, 181–186. (b)

Turnure, J. E. Reactions to physical and social distractors by moderately retarded institutionalized children. *Journal of Special Education*, 1970, **4**, 283–294. (c)

Turnure, J. E. Outerdirectedness in EMR boys and girls. *American Journal of Mental Deficiency*, 1973, **78**, 163–170.

Turnure, J., & Zigler, E. Outerdirectedness in the problem solving of normal and retarded children. *Journal of Abnormal and Social Psychology*, 1954, **59**, 427–436.

Yando, R., Zigler, E. Outerdirectedness in the problem-solving of institutionalized and noninstitutionalized normal and retarded children. *Developmental Psychology*, 1971, **4**, 277–288.

Zigler, E. Research on personality structure in the retardate. *International Review of Research in Mental Retardation*, 1966, **1**, 77–108.

Zigler, E., & Yando, R. Outerdirectedness and imitative behavior of institutionalized and noninstitutionalized younger and older children. *Child Development*, 1972, **43**, 413–425.

(Received February 18, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 926-931

TRAINING LETTER DISCRIMINATION BY PRESENTATION OF HIGH-CONFUSION VERSUS LOW-CONFUSION ALTERNATIVES[1]

ROSEMERY O. NELSON[2] AND KENNETH S. WEIN

University of North Carolina at Greensboro

Two groups of preschool children ($n = 8$ per group) were taught letter discriminations by means of a matching-to-sample task. The high-confusion group were presented with letter alternatives that were highly confusable, based on E. J. Gibson's reading research. The low-confusion group were presented with letter alternatives that were less likely to be confused. The performance on a posttest of letter discrimination of both training groups was superior to a no-treatment control group. While the high-confusion group required more training trials to criterion, its posttest performance was better than the low-confusion group ($p < .06$). These results were discussed in terms of a high-confusion training procedure producing better discriminations among distinctive features of letters.

Reading as an academic behavior has been considered theoretically within the framework of operant psychology by Bloom (1973), Goldiamond and Dyrud (1966), Skinner (1957), and Staats (1968). Although Gibson's theory of reading is usually placed within a cognitive framework (Williams, 1973), her theory is nonetheless amenable to experimentation employing an operant model of stimulus–response consequence. The present study is concerned with varying antecedent conditions to ascertain which were most likely to produce desirable reading responses.

Gibson's analysis of the reading process (Gibson, 1969) consisted of four phases: learning receptive and expressive language skills; learning to differentiate graphic symbols (i.e., letters); learning to decode letters into sounds (i.e., learning both variable and constant letter-sound relationship); and using progressively higher-order units of structure, including pronounceability, and semantic and syntactic constraints. Gibson's second phase, that of learning letter discrimination, is of primary interest in this study. Gibson held that letter discrimination is facilitated by recognition of the critical or distinctive features of letters. Her research (Gibson, Gibson, Pick, & Osser, 1962; Gibson, Osser, Schiff, & Smith, 1963) produced a set of stimulus materials, including both real letters and artificial letterlike forms (which can be categorized as highly confusable versus infrequently confusable) and also including a set of graphemic features that can be categorized as distinctive versus nondistinctive.

Four studies, using this set of stimulus materials, have concluded that the learning of distinctive features greatly facilitates letter discrimination. Pick (1965), using Gibson letterlike forms, taught three groups of kindergarten children to distinguish transformations from copies of three standard forms. Performance on a transfer task that involved different standard forms but

[1] This research was supported by the Research Council of the University of North Carolina at Greensboro, Grant 569. Portions of this paper were presented at the meeting of the Southeastern Psychological Association, Hollywood, Florida, May 1974.

[2] Requests for reprints should be sent to Rosemery O. Nelson, Department of Psychology, University of North Carolina, Greensboro, North Carolina 27412.

The authors are grateful for the assistance in this study offered by Leroy Delionbach, Linda Hay, Tommi Johnsen, and Arthur Peoples; and by the children and staff of Baynes' Happy Day Nursery, Busy Day Nursery, Cinder-Fella Pre-School, Hester's Creative Schools, and Randleman Park Day Care Center of United Day Care Services.

comparable transformations was superior to transfer tasks involving both different standards and different transformations, or the same standards and different transformations. Williams (1969), also using the Gibson letterlike forms with kindergarten children, employed three training methods during delayed matching to sample: discrimination training in which different forms were offered as alternative matches to the standard; discrimination training in which different transformations of the same form were offered as alternative matches to the standard form; and reproduction training in which the standard form was traced and copied. The middle group made the most errors during training, but performed significantly better than either of the other two groups on a transfer task also involving discrimination of letterlike forms. Tawney (1972) found that reinforcing four-year-old children to respond to critical dimensions of letterlike forms produced better performance on a matching-to-sample task involving real letters than reinforcing responses to noncritical dimensions. Samuels (1973) found that letter discrimination training that forced attention to distinctive features of letters facilitated learning letter names more than discrimination training not involving distinctive features or mere exposure to letters.

The present study differed from the first three summarized above in that real letters were used in both the training and transfer (posttest) tasks. The advantage of using real letters is that since these are the stimulus materials found in typical classrooms, the results of the study can be more readily generalized to the classroom. The hypothesis being tested, derived from the studies above, was that discrimination training involving highly confusable letters would produce better results on a transfer task (posttest) than discrimination training involving less confusable letters, because the former training would require greater attention to the distinctive features of the letters. The experimental design consisted of three groups of children: an experimental group taught to discriminate among low-confusion alternatives; an experimental group taught to discriminate among high-confusion alternatives; and a control group who were pre- and posttested with no additional training.

Method

Subjects

Fifty-three children, ages 2½ to 4½ years, who attended four local day care centers, completed a pretest of letter discrimination. Several other children were unable to complete pretesting due to crying or being unable to match colors during the familiarization slides. Of the 53 children, those who made at least 13 errors on the pretest consisting of 58 items were divided into three matched groups. Matching was done on the basis of age, sex, number of pretest correct responses, and the day care center which the child attended. For the high-confusion experimental group ($N = 8$: 5 females and 3 males), the mean age was 45.5 months, and the mean number of pretest correct responses was 40.125. For the low-confusion experimental group ($N = 8$: 5 females and 3 males), the mean age was 44.6 months, and the mean number of pretest correct responses was 40.500. For the control group ($N = 9$: 5 females and 4 males), the mean age was 43.3 months, and the mean number of pretest correct responses was 38.667. Two children were unable to complete the study: one child in the high-confusion experimental group cried during training, and one child in the control group became sick; these two children were replaced by two others so as to complete the subject pool described above.

Materials

Testing slides. The same set of slides was used during pre- and posttesting. There were 15 "familiarization" slides in which the child could either demonstrate his ability to match to sample or learn how to match to sample. These 15 familiarization slides consisted of squares of colors to be matched; a shaping procedure was used in that the first three slides had only two alternatives, the next four slides had three alternatives, the next four slides had four alternatives, and the final four slides had five alternatives. The actual pre- and posttest consisted of 58 slides containing two duplicate sets of 29 slides, in order to assess more reliably the child's skill in matching to sample by using letter stimuli. The slides were photographs of Roman capital letters. Each pre- and posttest letter slide consisted of one letter to be matched with five alternative choices in a horizontal row beneath it. The five alternatives, in random order from slide to slide, were the correct match, two low-confusion alternatives, and two high-confusion alternatives. A letter was judged to be highly confusable with another letter if they were confused four or more times in Gibson's Confusion Matrix I (Gibson et al., 1963). On the basis of this criterion,

six letters (A, F, J, P, S, and Z) did not appear as stimuli to be matched because they were not confused at least four times with two alternatives. On the other hand, if a letter was confused at least four times with three or four other letters, it became the stimulus in two different slides. A letter was judged to be of low confusion with another letter if both were confused no times in Gibson's Confusion Matrix I. Since there were many alternative low-confusion choices for any stimulus letter, a second criterion that had the least number of common distinctive features was employed, as determined by Gibson's chart of distinctive features for letters (Gibson, 1969, p. 88). As an example, M is highly confused with W and N, and rarely confused with L and Q.

Training slides. For each training group (high confusion and low confusion), there was also a set of 29 training slides using matching to sample with Roman capital letters. Each training slide consisted of the letter to be matched and three alternative choices in a horizontal row beneath it. For the high-confusion training group, the three alternatives (in random order from slide to slide) were the correct response and two high-confusion alternatives. For the low-confusion training group, the three alternatives (in random sequence from slide to slide) were the correct response and two low-confusion alternatives. Six color matching slides were also used as "warm-up" slides prior to each training session for both groups.

Apparatus. All slides were projected onto a wall by a Kodak carousel projector located a constant distance from the wall. Correct responses were followed by poker chips, which were used as tokens. The child deposited poker chips into a jar; all chips in the jar were exchanged at the end of the session for pieces of chocolate candy.

Procedure

Pretesting. Each subject was invited by a male examiner to participate in a "letter game." The child was taken from class in the day care center to a private room where the apparatus was set up. The child sat in a small chair facing the wall onto which the slides were projected. The child indicated a response by putting his or her finger on the wall on the choice of alternatives. During the color familiarization slides, a correction procedure was employed. If the child made an error, the correct response was pointed out, and he or she was permitted additional trials on the same slide. The few children who were unable to learn the matching-to-sample task during the familiarization slides were discontinued from the study at this point. The 58 pretest slides containing a stimulus letter and five alternatives were presented one time each without corrections; social reinforcement was used to maintain responding.

Training. Of the 53 children who completed pretesting, those who made 13 or more errors on the 58 slides were matched into one of three groups: high-confusion training, low-confusion training, or a control group who merely continued to attend the day care centers and was pre- and posttested without additional training. For low- and high-confusion training, each training session was conducted on an individual basis in the same room and using the same apparatus as during pretraining; the trainers were two female graduate students who trained the children on alternate days. For both low- and high-confusion training, each training session began with six warm-up slides involving matching to sample using colors. For the first training day, these warm-up slides were also used to teach the value of the tokens by having the child exchange tokens for candy at the end of these warm-up slides. Each training session consisted of completing the 29 slides appropriate to that training group. If the child made a correct response, he or she was given a token and social reinforcement, and the next slide appeared. If the child made an incorrect response, the experimenter said "no," and the slide projector was turned off for five seconds (time out); the same slide then reappeared. This procedure was followed until the child made the correct response for that slide. The trainer recorded all incorrect responses on precoded data sheets. Each session ended when the child had correctly matched the letters on each of the 29 training slides. The child was then allowed to exchange tokens for chocolate candy. The sequence of slides for both the high- and low-confusion training groups were randomized daily. A child completed the entire training sequence when he or she had two consecutive sessions in which no errors were made, or when a maximum of 20 sessions had been completed. No child needed more than 17 training sessions to meet the former criterion.

Posttesting. Posttesting was conducted within three days of the time when the child completed the last training session. For every two experimental children who completed training and posttesting, one control child was also posttested. The same slides were used during posttesting as during pretesting; these slides consisted of familiarization slides and 58 slides consisting of a stimulus letter and five alternative choices from which to select the correct match. The same male examiner who conducted pretesting carried out the posttesting; he remained "blind" as to which group the child had been assigned (high confusion, low confusion, or control).

Results

Trials to Criterion

The criterion for completion of training for each child in the high-confusion and low-confusion experimental groups was two consecutive days with no errors on the set of 29 training slides appropriate for each training group. The mean trials to criterion (num-

TABLE 1

MEAN NUMBER OF CORRECT RESPONSES, ERRORS, AND TRIALS TO CRITERION

Group	N	Pretest				Posttest				Trials to criterion	
		Mean score[a]	*SD*	HC errors	LC errors	Mean score[a]	*SD*	HC errors	LC errors	*M*	*SD*
HC	8	40.125	3.68	17.250	.625	54.000	3.78	3.750	.250	6.625	4.44
LC	8	40.500	8.18	15.875	1.625	49.125	5.25	8.750	.125	2.750	1.39
C	9	38.667	4.56	18.778	.617	41.778	7.47	15.111	1.111	not trained	not trained

Note. HC = high confusion, LC = low confusion, C = control.
[a] Maximum score = 58.

ber of training days required) for the high-confusion group was 6.625; for the low-confusion group, 2.750. A t test was performed on these mean differences, which revealed that significantly more training days ($t = 2.357$, $df = 14$, $p < .05$) were required for the high-confusion group to meet criterion than for the low-confusion group.

Pretest-to-Posttest Difference Scores

All subjects were given the same pre- and posttest of letter discrimination consisting of 58 matching-to-sample slides, each with one stimulus letter and five alternative letters (the match, two high-confusion letters, and two low-confusion letters). All groups improved from pre- to posttest. The mean pretest, posttest, and difference scores for all three groups are summarized in Table 1. A one-way analysis of variance revealed significant differences among the pretest-to-posttest difference scores for the three groups ($F = 8.99$, $df = 2.22$, $p < .01$). A Newman-Keuls test for multiple comparisons among means demonstrated that the high-confusion group was different from the control group at the .01 level of significance; that the low-confusion group was different from the control group at the .05 level; and that the high-confusion group was different from the low-confusion group approaching the .05 level of significance ($p < .06$). Thus both training groups performed significantly better on the posttest (as compared with their pretest performances) than the no-treatment control group. While the high-confusion group required significantly more training trials than the low-confusion group, their performance was significantly better ($p < .06$) on the posttest of letter discrimination than the low-confusion group.

Error Analysis

An examination of Table 1 reveals that for all three groups, the most frequent type of error made on the pretest was errors between high-confusion letters. The criterion for a letter to be judged highly confusable with another letter, it will be recalled, is four or more errors in confusion between them in Gibson's Confusion Matrix I. Since during training, the low-confusion group was discriminating only among low-confusion alternatives, and the high-confusion group was discriminating only among high-confusion alternatives, an analysis of the types of error made during the posttest for these two training groups was undertaken. The means reported in Table 1 indicate that regardless of type of training procedure, more than 90% of posttest errors for all three groups were confusions between high-confusion letters.

Confusion Matrix

In order to corroborate Gibson's Confusion Matrix I (Gibson et al., 1963), the pretest errors of all 53 children who completed pretesting were analyzed. Given the construction of the pretest, unlike Gibson's matrix, each letter did not have an equal opportunity to be confused with every other letter. Hence the number of opportunities for such confusions was tabulated for each letter-pair combination. A percentage measure was taken by dividing the actual number of confusions for each letter pair by the number of opportunities for such confusions.

The pairs of data in 77 corresponding cells produced a Pearson correlation of .783 between the confusion matrix derived from the pretest data of the present study and Gibson's confusion matrix.

Discussion

This study indicated that training in letter discrimination resulted in better posttest performance than lack of such training. While subjects who were trained by making discriminations among high-confusion letters required more trials to reach criterion, their posttest performance was better ($p < .06$) than subjects who were trained by making discriminations among low-confusion letters.

These results corroborate those of Pick (1965), Williams (1969), Tawney (1972), and Samuels (1973), each of whom concluded that letter (letterlike) discrimination was enhanced by teaching children to attend to the critical features of the stimuli. Of these studies, Williams' and Samuels' are most similar to the present investigation; both studies indicate that although training with frequently confused graphic stimuli results in slower acquisition, this training produces better performance on a transfer task than training with infrequently confused stimuli.

The fact that discrimination learning requires differential training on two or more stimuli is well accepted (e.g., Lashley & Wade, 1946). In animal operant research, it has been concluded that the difficulty of the discrimination is inversely related to the difference between the two training stimuli. The generalization gradient, however, is steeper if the two stimuli that are used in training are closely related on a stimulus dimension (Terrace, 1966). The teaching of letter discrimination by differential training using high-confusion alternatives would be predicted to produce many training trials but a better performance on further tests involving letter discrimination; these were indeed the obtained results.

Similar results have been obtained from discrimination training involving words as another type of verbal stimuli. Using lists of words, Levin and Watson (1963), McCutcheon and McDowell (1969), Otto and Pizzillo (1970), and Samuels and Jeffrey (1966) all conclude that high intralist similarity impedes rate of learning of the original list, but produces greater accuracy on a transfer list of words.

Guralnick (1972) reviewed the research available on letter discrimination and concluded that the educational implication of this research was that pretraining on the distinctive features of letters would facilitate reading acquisition. The present study provides a specific suggestion to the classroom teacher: Letter discrimination learning is facilitated by training the distinctions between two similar letters. While using two dissimilar letters may result in faster acquisition, this is what Samuels and Jeffrey (1966) labeled "false economy," since poorer results are produced in transfer situations.

One of the questions that remains to be answered is whether providing the low-confusion group with overlearning trials—so that their total number of discrimination learning trials is yoked to that of the high-confusion group—would change the present results. Although the present design was analogous to classroom stiuations in which teaching ceases with mastery, some possibility remains that the superior posttest performance of the high-confusion group was not due to the nature of the training materials but rather simply to the greater number of training trials that this group underwent.

In addition to demonstrating the posttest superiority resulting from high-confusion training, the second main conclusion from the present study was a corroboration of Gibson's Confusion Matrix I (Gibson et al., 1963), since the letter-confusion errors of the present study correlate with Gibson's Confusion Matrix I with r equal to .783. Gibson et al. (1962) had found a correlation of .87 between confusion errors made with letterlike forms and confusion errors made with real letters. Since these letterlike forms were constructed from Gibson's theoretical chart of distinctive features of graphemes (letters), the present study indirectly through this series of correlations supports the use of Gibson's stimulus materials in

future research as samples of high-confusion versus low-confusion letters or letterlike forms, or as distinctive versus nondistinctive graphemic features.

REFERENCES

Bloom, R. Learning to read: An operant perspective. *Reading Research Quarterly*, 1973, **8**, 147–166.

Gibson, E. J. *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts, 1969.

Gibson, E. J., Gibson, J. J., Pick, A. D., & Osser, H. A developmental study of the discrimination of letter-like forms. *Journal of Comparative and Physiological Psychology*, 1962, **55**, 897–906.

Gibson, E. J., Osser, H., Schiff, W., & Smith, J. An analysis of critical features of letters, tested by a confusion matrix. In, *A basic research program on reading*. (Cooperative Research Project No. 639) Washington, D.C.: United States Office of Education, 1963. (ERIC Document Reproduction Service No. 002967)

Goldiamond, I., & Dyrud, J. C. Reading as operant behavior. In J. Money (Ed.), *The disabled reader*. Baltimore: Johns Hopkins Press, 1966.

Guralnick, M. J. Alphabet discrimination and distinctive features: Research review and educational implications. *Journal of Learning Disabilities*, 1972, **5**, 427–434.

Lashley, K. S., & Wade, M. The Pavlovian theory of generalization. *Psychological Review*, 1946, **53**, 72–87.

Levin, H., & Watson, J. The learning of variable grapheme-to-phoneme correspondences. (Final Rep.) Washington, D.C.: Basic Research Program on Reading, United States Office of Education, 1963. (ERIC Document Reproduction Service No. 002967)

McCutcheon, B. A., & McDowell, E. E. Intralist similarity and acquisition and generalization of word recognition. *Reading Teacher*. 1969, **23**, 103–107, 115.

Otto, W., & Pizzillo, C. Effect of intralist similarity in kindergarten pupils' rate of word recognition and transfer. Paper presented at the meeting of the American Educational Research Association, New York, February 1970. (ERIC Document Reproduction Service No. 047895)

Pick, A. D. Improvement of visual and tactual form discrimination. *Journal of Experimental Psychology*, 1965, **69**, 331–339.

Samuels, S. J. Effect of distinctive feature training on paired-associate learning. *Journal of Educational Psychology*, 1973, **64**, 164–170.

Samuels, S. J., & Jeffrey, W. E. Discriminability of words and letter cues used in learning to read. *Journal of Educational Psychology*, 1966, **57**, 337–340.

Skinner, B. F. *Verbal behavior*. New York: Appeton-Century-Crofts, 1957.

Staats, A. W. *Learning, language, and cognition*. New York: Holt, Rinehart & Winston, 1968.

Tawney, J. W. Training letter discrimination in four-year-old children. *Journal of Applied Behavior Analysis*, 1972, **5**, 455–465.

Terrace, H. S. Stimulus control. In W. K. Honig (Ed.), *Operant behavior: Areas of research and application*. New York: Appleton-Century-Crofts, 1966.

Williams, J. P. Training kindergarten children to discriminate letter-like forms. *American Educational Research Journal*, 1969, **6**, 501–514.

Williams, J. P. Learning to read: A review of theories and models. *Reading Research Quarterly*, 1973, **8**, 121–146.

(Received March 4, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 932-939

MULTITRAIT-MULTIMETHOD ANALYSIS OF CONCEPTUAL TEMPO

VERNON C. HALL[1] AND WILLIAM J. C. RUSSELL

Syracuse University

The present study used the multitrait-multimethod matrix to determine the convergent and divergent validity of conceptual tempo. Two measures of conceptual tempo (Matching Familiar Figures Test, MFF, and Word Recognition Test, WRT) and two measures of intellectual functioning (Raven Coloured Progressive Matrices, RCPM, and Peabody Picture Vocabulary Test, PPVT) were presented to 82 third-grade boys twice (to acquire reliabilities). It was found that the latencies on all instruments correlated with latencies on all other instruments (no divergent validity) and that error scores on the MFF correlated better with error scores on the RCPM than the WRT.

In 1964 Kagan, Rosman, Day, Albert, and Phillips reported a series of experiments on information processing. While investigating conceptual style, that is, whether a child associates objects or pictures of objects analytically or relationally, Kagan noted that children who reported more analytic concepts had a tendency to delay their response longer than children who reported more relational concepts. This led to the hypothesis of a more fundamental construct, which was responsible for reflective children producing more relational responses. Confirmation for this hypothesis was found in the positive relationship between the number of correct solutions on the first response and the amount of latency to the first response on a Design Recall Test (DRT) constructed by Kagan. However, it was felt that the DRT might contain a memory factor that was a part of verbal intelligence. Because Kagan felt measures of this hypothesized construct were orthogonal to intelligence, he constructed the Matching Familiar Figures Test (MFF) in an effort to provide a more appropriate (pure) measure of the construct. This MFF (briefly described later) has become the most used measure of the construct Kagan has labeled "conceptual tempo."

[1] Requests for reprints should be sent to Vernon C. Hall, Syracuse University, 331 Huntington Hall, Syracuse, New York 13210.

Based on the hypothesis that time used for reflection and analysis leads to fewer errors and that fast responding leads to many errors in a standard variant task requiring response uncertainty, significant negative correlations between latencies and errors on the MFF should be expected. In 14 studies reviewed by the present authors there were 37 negative correlations significant at the .01 level, 4 significant at the .05 level, and 5 failed to reach significance. Only one nonsignificant positive correlation was found.

Using a variety of perceptual recognition tasks, Kagan et al. (1964) found conceptual tempo (latency) to be stable over time. "The reliability of response times over a nine-week period was in the high 70s and there was remarkable generality across tasks as varied as DRT, MFF and HVM [p. 32]."

Thus with relatively high negative correlations between errors and latencies on the MFF; high positive correlations between the MFF and similar perceptual tasks involving high response uncertainty and simultaneous presentation of alternative hypotheses, and adequate reliability on test-retest correlations, the MFF has been accepted as an appropriate tool for classifying the response patterns of lower elementary school subjects (Grades 1–4) using the con-

ceptual tempo construct as either impulsive or reflective.

Subsequently, Kagan has devised a classification system, employing the two measures (latency to first response and number of errors), that uses a double median split. Those subjects scoring below the median on errors and above the median on latency are classified as reflective; that is, they take longer to respond, but their first response is usually correct. Those above the median on errors and below the median on latency are classified as impulsive; that is, they make fast initial response, but that response is often incorrect. This classification has typically resulted in categorizing approximately 70% of the sample as either reflective or impulsive with the remaining 30% distributed about equally between the fast-correct and slow-wrong groups.

In most of the subsequent research, conceptual tempo has been widely accepted and studied as an important new dimension of intellectual development independent from intelligence. It is obvious that if this is indeed a new trait, it is an important one for education. Teachers, unaware of this native response tendency would tend to make school a relatively unhappy place for children by unnecessarily and inappropriately hurrying the reflectives or slowing the impulsives.

It is important in psychology, however, to examine carefully the possibility that any new trait is indeed new rather than simply being a different term for something already known. Campbell and Fiske (1959) recognized this danger and pointed out that measures of individual differences have typically been validated by the means of convergence; that is, a significant correlation between a new measure of a concept and an independent and accepted measure of the same concept has been a common procedure for validating a new construct (e.g., MFF and DRT). They go on to suggest that a new construct should also demonstrate discriminant validity. That is, a test measuring a construct can be invalidated by too high a correlation with another test from which it is intended to differ (because it measures a different construct).

In the present case the authors were concerned that it be demonstrated more conclusively that the conceptual tempo trait has discriminant validity with the trait of intelligence. A review of the literature led to the conclusion that the evidence for discriminant validity is less than conclusive. From observing Table 1, it can be determined that of the 42 correlations reported between MFF errors and measures of intellectual functioning in 11 studies, 21 were significant. Fourteen of 42 correlations between latencies on the MFF and measures of intellectual functioning were significant.

In order to examine both the convergent and discriminant validity of a trait, Campbell and Fiske (1959) suggested that more than one trait be measured and that more than one method be used to measure each trait. Results can then be plotted in a multitrait-method matrix and examined for reliability (the same trait by test–retest), convergent validity (the same trait measured by different methods), and discriminant validity (different traits measured by the same method). In the latter case, relatively low correlations due to common method variance would be expected.

The purpose of the present study was to employ the multitrait-multimethod matrix to determine whether the conceptual tempo individual difference construct is (*a*) stable over time, (*b*) stable across instruments, and (*c*) orthogonal with measures of intellectual functioning.

Method

Subjects

Subjects for this study were 85 white, third-grade boys from six classrooms in two elementary schools in a middle-class suburban school district (3 subjects were dropped because they were absent on their scheduled date for retest; thus the analyses contained 82 subjects). The mean age of the subjects was 8 years, 10 months, with a standard deviation of 6.5 months.

Boys were selected because most studies using both sexes have found significant sex differences in responses on the MFF. Third-graders were used because Kagan et al. (1964) developed the MFF while studying children in Grades 1 through 4. Achenbach (1970) suggested that the MFF decreases in validity with fifth- and sixth-graders.

TABLE 1

CORRELATION BETWEEN TIME AND ERROR SCORES ON THE MATCHING FAMILIAR FIGURES TEST AND SCORES ON OTHER MEASURES OF INTELLIGENCE AND ACHIEVEMENT

Study	Errors—IQ		Latency—IQ		Grade	Instrument
	Male	Female	Male	Female		
Kagan, Rosman, Day, Albert, & Phillips, 1964	−.53**	−.28	.36	.22	3	IQ based on WISC, Verbal scale IQ
	−.47**	−.21	.05	.15	3	IQ based on CTMM
	.10	−.40*	−.13	−.03	4	IQ based on WISC, Verbal scale IQ
Kagan, 1965a	−.26	−.32	.27	.35**	1	IQ based on WISC, Verbal scale IQ
	−.25	−.20	.00	.46**	2	IQ based on WISC, Verbal scale IQ
	−.36**	−.30	.36**	.19	3	IQ based on WISC, Verbal scale IQ
	−.38**	−.02	.12	−.16	3	IQ based on WISC, Verbal scale IQ
Kagan, 1966a	−.29**	−.20	.08	.25*	1	IQ based on WISC, Verbal scale IQ
Kagan, 1966b	−.53**	−.28	.36**	.22	2, 3	IQ based on WISC, Verbal scale IQ
	.10	−.40*	−.13	−.03	3, 4	Same subjects retested
Lewis, Rausch, Goldberg, & Dodd, 1968	−.30	−.60**	.30	.45*	Preschool	IQ based on Stanford-Binet
	−.40	−.67**				
Yando & Kagan, 1968	−.36**	−.33**	.20	.27*	1	Metropolitan Reading Readiness Test
	−.23*	−.48**	.16	.40**	1	Same subjects retested
Ward, 1968	−.09	−.33**	.23	.23	Kindergarten	IQ based on PPVT
	−.27	−.22	.15	.39**	Kindergarten	IQ based on PPVT
Meichenbaum & Goodman, 1969	.62**		.39*		Kindergarten	Thurstone PMA Test
Eska & Black, 1971	−.47**	−.35*	.33*	.45**	3	IQ based on Otis-Lennon
Harrison & Nadelman, 1972	−.30	−.34*	.46*	.00	Preschool	IQ based on PPVT
	−.29*		.26			
Lindstrom, 1972	.03		.02		2	WISC, verbal scale IQ
	−.05		−.03		2	WISC, performance scale IQ
	−.02		.00		2	WISC, full scale IQ
	.45**		−.33**		2	Raven Colored Progressive Matrices (errors)

Note. WISC = Wechsler Intelligence Scale for Children; CTMM = California Test of Mental Maturity; PPVT = Peabody Picture Vocabulary Test; PMA = primary mental abilities.
* $p < .05$.
** $p < .01$.

Thus, third-graders should be experienced enough in the school setting to feel secure in the testing situation and well within the validity range of the MFF.

Instruments.

Four instruments were employed in this study, two designed to measure conceptual tempo and two

designed to measure intellectual functioning. All shared the same format (multiple choice) and yielded the same scores (latency, number correct, and errors).

Measures of conceptual tempo were as follows:

1. *Matching Familiar Figures Test.* As mentioned earlier this test has been the most frequently used instrument to measure conceptual tempo. This 12-item test (there are now several forms but the original was used), a visual matching task, requires the subjects to select the one stimulus from an array of six variants that is identical to a standard. The stimuli consist of line drawings of familiar objects.

2. *Word Recognition Test* (WRT). This 20-item test, which was constructed by the present experimenters, was patterned after a test developed by Kagan (1965b) to study word-recognition skills. He found, "Word-recognition errors were negatively related to response time on MFF for both sexes (coefficients were significant at .01 or beyond for the two-tailed test). In addition, high-error scores on MFF and HVM each predicted high word-error scores ($p < .05$ or beyond) [Kagan, 1965b, p. 615]."

For this test, two of Kagan's test items (included in the original article) were used as practice items. Included in the 20 test items were four types of stimuli: (*a*) Words in which the initial phonemes were similar; that is, *shone, shore, shove, shame,* and *shave,* with *shove* being the stimulus word. (*b*) Words in which the final phonemes were identical; that is, *bang, rang, hang, sang, gang,* with *rang* being the stimulus word. (*c*) Multisyllable words with the same initial phoneme; that is, *quagmire, quadrang, quadrate, quartile,* and *quadroon,* with *quadroon* being the stimulus word. (*d*) Multisyllable words with the same final phoneme; that is, *reflective, additive, sensitive, tentative,* and *creative,* with *sensitive* being the stimulus word.

All positions were represented an equal number of times for right answers. As an additional attempt to ensure response uncertainty, all words were on or above the third-grade reading level. For both measures of conceptual tempo the items were scored for response time to first selection, number of errors before finding the correct response, and number of correct first responses.

Measures of intellectual functioning were as follows:

1. *Raven's Coloured Progressive Matrices* (RCPM, 1965). In this 36-item test, the subject is asked to select the correct variant from an array of six variants that will complete a matrix. In a review of more than 70 studies, Burke (1958) concluded that while the RCPM was not a substitute for the Stanford Binet or Wechsler tests, it "is perhaps an almost equally useful supplement and shows intercorrelations with such tests perhaps as high as they show with one another [p. 222]." Burke found reports of studies using normal children in which there were correlations between the RCPM and the Wechsler Intelligence Scale for Children (WISC) full-scale, verbal, and performance IQs of .91, .84, and .83, respectively.

2. *Peabody Picture Vocabulary Test* (PPVT). In this test the child is asked to select the correct picture from an array of four pictures that represent the stimulus word pronounced by the experimenter. The PPVT was developed as a measure of verbal intelligence, and the manual reports a number of high correlations with other standardized measures of intellectual ability. For both measures of intellectual functioning, items were scored for response time to first selection, number of errors, and number of correct responses. Because these two tasks did not require that the child continue to guess until the correct answer was achieved (except for the first two items on the RCPM), the error scores were mirror images of the correct scores.

Procedures

Subjects were tested individually in an empty room. All subjects were eager to participate on both the test and retest. To control for order effects, each order was administered an equal number of times (seven), with the exception of four orders that were administered six times. In addition, for each subject, the test and retest orders were different.

A test–retest design, with a one-week interval between test sessions, was used to secure reliabilities for the multimethod-multitrait correlational analysis. Standardized instructions already available were used for all instruments except the WRT where the following instructions were used.

"In a minute I will say a word, then I will show you a card with five words printed on it. I want you to point to the word I said. If you point to the wrong one, I will just say 'no, try again' until you get the right one."

Analyses

Correlations between methods and traits were computed and analyzed for convergent and discriminant validity following the suggestions of Campbell and Fiske (1959). Three correlation matrices were computed; one using results from first administration, one using results from the second administration, and one in which each pair of correlations was transformed into a mean correlation, using Fisher's table for r to z transformations.

As an additional estimate of the stability of the MFF as a measure of conceptual tempo, Cramer's coefficient (Hays, 1965) was calculated on the dispersion of subjects from the diagonal of a fourfold table (see Table 3). If a subject obtained the same classification, he would be classified in one of the cells on the diagonal of the fourfold table. This statistic was computed for each instrument.

TABLE 2
MEANS AND STANDARD DEVIATIONS FOR ALL INSTRUMENTS

Instrument	Error		Correct		Latency	
	Test	Retest	Test	Retest	Test	Retest
MFF						
M	16.35	15.38	3.87	4.44	8.51	7.30
SD	5.02	5.57	1.77	1.73	3.95	4.78
WRT						
M	3.41	2.72	17.95	17.89	3.38	2.87
SD	2.63	3.33	2.01	2.25	1.28	1.56
RCPM						
M	14.99	12.66	21.57	23.82	4.57	3.76
SD	4.46	4.68	4.51	4.70	1.41	1.10
PPVT						
M	49.66	48.32	75.34	76.68	3.27	2.53
SD	8.48	8.10	8.48	8.10	1.29	1.08
PPVT IQ						
M			108.41	110.57		
SD			15.03	14.23		

Note. MFF = Matching Familiar Figures Test; WRT = Word-Recognition Test; RCPM = Raven Coloured Progressive Matrices; PPVT = Peabody Picture Vocabulary Test. $N = 82$.

RESULTS

Table 2 includes the means and standard deviations for each instrument on the test and retest. It can be seen from the means and standard deviations of IQ scores on the PPVT that these subjects were slightly above average but well within the normal range.

Reliability

The values on the diagonal of Table 4 are the reliabilities. It can be seen that all of the reliabilities are significant at the .05 level and all except one (MFF correct score) are significant at the .01 level. Interestingly enough, two of the three lowest values are for the MFF. Table 3 includes the frequency distributions and Cramer's coefficients for the four measures. It should be noted that while all of the values are significant (meaning there was stability across administrations), the distribution of the MFF scores more closely resembled that of the RCPM than the WRT.

Convergent Validity

Table 4 also contains the mean correlation values acquired from Fisher's r to z transformation, which the authors believe is the most appropriate for purposes of the present study. The only correlations that might have led to different conclusions (i.e., were significant on one computation but not another) were the correlations between MFF correct and RCPM correct, PPVT errors and RCPM errors, and MFF errors and PPVT errors, which were significant on the retest only. Convergent validity was determined by the relationship between the same measures (error, correct, and latency) derived from two instruments designed to measure the same traits (MFF and WRT for conceptual tempo; RCPM and PPVT for intellectual ability). The relevant correlations are underlined once in Table 4. Although few of the correlations are particularly impressive, all three scores acquired from the two measures of intellectual ability are significantly correlated while only the latency score is significantly correlated for the two measures of conceptual tempo.

Discriminant Validity

The discriminant validity was ascertained by determining the correlations between the same scores obtained on the tests of different traits. The relevant correlations are underlined twice in Table 4. In this case the three scores on the two measures of con-

ceptual tempo were significantly correlated with the three scores obtained on the RCPM (a different trait). These correlations were all higher than the correlations representing convergent validity for the conceptual tempo trait. In addition, the latency score (which Kagan maintains in the most accurate measure of conceptual tempo) on both measures of conceptual tempo are correlated with the scores obtained on the PPVT.

Discussion

The authors believe that these results add important information concerning the nature of what has been labeled conceptual tempo. In the first place, there does indeed seem to be a consistent tendency to display slow or fast response times in problem situations. The latency reliabilities were all quite high, and latencies for all measures were significantly correlated with latencies for all other measures. It appears that if one wants to use response latency as a measure of conceptual tempo, then any of these instruments would suffice. In effect, there was no divergent validity demonstrated for the latency measure.

The degree to which response uncertainty is a necessary condition for this tendency to be displayed is somewhat less clear. If one maintains that a negative correlation between response time and errors is necessary evidence for response uncertainty because quick responders will be inaccurate, then the answer is no. On both the WRT and PPVT the relationship between errors and latency was nonsignificant (yet the latency reliability for the WRT was the highest of all measures). This would mean that the words used on the WRT were not enough alike to produce sufficient uncertainty for these children.

When we turn to errors or number correct, the picture changes. The reliability for these measures on the MFF is the lowest of the four instruments used. This may be due to the fact that it is really a learning task (e.g., subjects keep choosing until they locate the correct answer). On the other hand, the mean improvement on the MFF from test to retest was less than one item

TABLE 3

Test–Retest Distributions and Cramer's Coefficients for the Four Instruments Using the Double Median Split

Post test	Pretest			
	Slow and wrong	Reflective	Fast and right	Impulsive
Matching Familiar Figures Test (ρ' = .411)				
Slow and wrong	3	6	1	7
Reflective	6	14	2	3
Fast and right	3	1	1	7
Impulsive	5	3	7	13
Raven Coloured Progressive Matrices (ρ' = .437)				
Slow and wrong	9	2	0	4
Reflective	5	12	7	2
Fast and right	2	6	2	2
Impulsive	4	1	3	21
Word Recognition Test (ρ' = .492)				
Slow and wrong	18	6	0	5
Reflective	4	8	3	0
Fast and right	0	0	16	6
Impulsive	3	2	5	6
Peabody Picture Vocabulary Test (ρ' = .593)				
Slow and wrong	16	2	1	4
Reflective	6	10	2	0
Fast and right	0	3	18	2
Impulsive	2	2	2	12

(the improvement on both the PPVT and RCPM was more than one). In addition, the same reliabilities for the WRT (presented in the same manner as the MFF) were excellent. This low reliability brings into question the procedure of using a double median split to classify subjects as reflectives and impulsives (over half of the students changed in classification from pretest to posttest).

The contention that conceptual tempo is independent from verbal intelligence is supported by the lack of significant relationship between errors and number correct on the MFF and the same measures on the PPVT.

The fact that the latencies were significally correlated with number correct (and number wrong) on both the MFF and RCPM, and that there were also significant

TABLE 4

MEANS OF THE TEST AND RETEST CORRELATIONS, ARRANGED ACCORDING TO METHOD

Method	1	2	3	4	5	6	7	8	9	10	11	12
CT_1 (MFF)												
1. Correct	.260											
2. Error	−.788	.340										
3. Latency	.520	−.540	.680									
CT_2 (WRT)												
4. Correct	.108	−.146	.250	.710								
5. Error	−.090	.126	−.235	−.960	.730							
6. Latency	.126	−.106	.452	−.115	.110	.810						
IQ_1 (RCPM)												
7. Correct	.238	−.310	.386	.376	−.346	.110	.750					
8. Error	−.250	.340	−.404	−.376	.330	−.106	−.970	.720				
9. Latency	.212	−.206	.536	.166	−.110	.586	.475	−.495	.660			
IQ_2 (PPVT)												
10. Correct	.090	−.186	−.090	.126	−.076	−.080	.266	−.230	.045	.830		
11. Error	−.090	.186	−.090	−.126	.076	.080	−.266	.230	−.045	−1.000	.830	
12. Latency	.182	−.192	.405	.156	−.120	.400	.366	−.262	.597	−.015	.015	.470

Note. CT = conceptual tempo; MFF = Matching Familiar Figures Test; WRT = Word-Recognition Test; RCPM = Raven Coloured Progressive Matrices; PPVT = Peabody Picture Vocabulary Test, $r \geq .220$, $\alpha = .05$; $r \geq .286$, $\alpha = .01$. Single underlined values are relevant for convergent validity. Double underlined values are relevant for discriminant validity.

correlations between errors and number correct on the MFF with errors and number correct on the RCPM, gives us some indication of the kinds of abilities measured by the MFF. Here one could contend that the MFF is a rather poor (due to the low test–retest reliability) measure of the same abilities as those measured by the RCPM. These abilities have recently beome the subject of considerable interest in developmental and educational psychology.

REFERENCES

Achenbach, T. M., The children's associative responding test: A possible alternative to group IQ tests. *Journal of Educational Psychology,* 1970, **61,** 340.

Burke, H. R., Raven's Progressive Matrices: A review and critical evaluation. *The Journal of Genetic Psychology,* 1958, **93,** 199–228.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, **56,** 81–105.

Eske, B., & Black, K. N. Conceptual tempo in young grade-school children. *Child Development,* 1971, **42,** 505–516.

Harrison, A., & Nadelman, L. Conceptual tempo and inhibition of movement in black preschool children. *Child Development,* 1972, **43,** 657–668.

Hays, W. L. *Statistics for psychologists,* New York, Holt, Rinehart and Winston, 1965.

Kagan, J. Impulsive and reflective children. In J. Krumboltz (Ed.), *Learning and the educational process.* Chicago: Rand McNally, 1965. (a)

Kagan, J. Reflection-impulsivity and reading ability in primary grade children. *Child Development,* 1965, **36,** 609. (b)

Kagan, J. A developmental approach to conceptual growth. *Analyses of concept learning.* New York: Academic Press, 1966. (a)

Kagan, J. Developmental studies in reflection and analysis. In A. H. Kidd & J. H. Rivoire (Eds.), *Perceptual development in children.* New York: International Universities Press, 1966. (b)

Kagan, J., Pearson, L., & Welch, J. Conceptual impulsivity and inductive reasoning. *Child Development,* 1966, **37,** 593–594. (a)

Kagan, J., Pearson, L., & Welch, J. Modifiability of an impulsive tempo. *Journal of Educational Psychology,* 1966, **57,** 359–365. (b)

Kagan, J., Rosman, B. L., Day, D., Albert, J., & Phillips, W. Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs,* 1964, **78**(No. 1, Whole No. 578).

Lewis, M., Rausch, M., Goldberg, S., & Dodd, C. Error, response time, and IQ: Sex differences in cognitive style of pre-school children. *Perceptual and Motor Skills,* 1968, **26,** 563–568.

Lindstrom, D. Styles of cognitive behavior in second grade males: The relationship between conceptual tempo and responsiveness to cognitive demand. Unpublished manuscript, Syracuse University, 1972.

Meichenbaum, D., & Goodman, J. Reflection-impulsivity and verbal control of motor behavior. *Child Development,* 1969, **40,** 785–797.

Raven, J. C. *Guide to using the coloured progressive matrices: Sets A, Ab, B.* London: H. K. Lewis, 1965.

Ward, W. C. Reflection-impulsivity in kindergarten children. *Child Development,* 1968, **39,** 867–874.

Yando, R. M., & Kagan, J. The effect of teacher tempo on the child. *Child Development,* 1968, **39,** 27–34.

(Received March 5, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 940-947

EFFECTS OF VICARIOUS CONSEQUENCES AND MODEL AND EXPERIMENTER SEX ON IMITATIVE BEHAVIOR IN FIRST-GRADE CHILDREN

CANDACE S. GARRETT[1] AND DONALD J. CUNNINGHAM

Indiana University

This study investigated the effects of sex of model and experimenter and the consequences (verbal reward, punishment, or ignoring) of the model's behavior in a non-sex-typed modeling situation on the imitative behavior of first-grade children. The results indicated (*a*) reward and ignore conditions were not different but both yielded higher imitative scores than the punish condition; (*b*) same-sex models (reward condition) yielded higher imitation scores than opposite-sex models, (*c*) highest imitation scores (reward condition) were obtained by children exposed to same-sex models and experimenters, and lowest imitation scores (punish condition) by children exposed to a male experimenter and a female model. These results are discussed in relation to a social learning theory of sex-role development and methodological considerations.

Current popular interests in maximizing children's potentials and in guaranteeing individuals' rights have raised concerns that some sex-typed behaviors and attitudes may be potentially limiting to children of both sexes. An understanding of the development of these behaviors is an important and necessary first step toward revising, if possible, the environmental constraints that produce these limitations. A survey of the available literature (Garrett, Cunningham, & Buelow, 1974) unearthed only a few experimental studies concerning aspects of sex-role development. Most of the research reviewed in that paper came from studies that were not designed to investigate the effects of sex in the various experimental situations; sex was usually secondary to the purposes of the studies. Moreover, the theoretical issues examined in those papers were not usually tied directly to sex-role development. There is, therefore, a need for theoretically based empirical research aimed directly at understanding the course of sex-role development.

Three major sets of theories can be identified that purport to describe and explain sex-role development: psychoanalytic, cognitive-developmental, and social learning theory (Mischel, 1970). The last theory appears to be the most appealing as it is both comprehensive and amenable to operational formulation. According to social learning theory, the development and emergence of sex-typed behaviors and attitudes can be described by the same learning principles used to account for any other aspect of a person's social behavior, generally principles related to reinforcement and imitation (Bandura, 1971). This theory emphasizes characteristics of the subject and his or her environment that contribute to and affect sex-role development.

The results of past research related to sex-role development correspond reasonably well with this theory. Also, social learning theory has been systematically investigated as a basic model for a wide variety of social-emotional learning, such as aggression (Bandura, Ross, & Ross, 1963a), physical affectionate behavior (Fryrear & Thelen, 1969),

[1] Requests for reprints and for copies of the original technical report should be sent to Candace S. Garrett, Institute for Child Study, 2805 East Tenth Street, Indiana University, Bloomington, Indiana 47401.

self-reward criteria (Bandura & Kupers, 1964), and moral judgments (Bandura & McDonald, 1963). If the acquisition of sex-role behaviors and attitudes can be analyzed in terms of this learning theory and if the relevant subject and environmental variables can be identified and interrelated, then it should be possible to foster less stereotyped, more flexible, sex-role adaptions through environmental manipulations.

Of interest in this study were the following variables: vicarious reward or punishment, the sex of the subject, the current sex-role preference of the subject, the sex of the model, and the sex of the experimenter who managed the imitation situation and dispensed verbal rewards or punishments to the model as a consequence of the model's behavior. With respect to vicarious reward or punishment, it was expected that children—when placed in an imitation situation where they observe a model's responses either rewarded, punished, or ignored—will imitate models who are verbally rewarded, counterimitate (i.e., choose the opposite responses from) models who are verbally punished, and neither imitate nor counterimitate models who are verbally ignored. Support for this hypothesis can be found in Bandura, Ross, and Ross (1963b) and Liebert, Sobol, and Copemann (1972), among others (see Garrett, Cunningham, & Buelow, 1974, for further discussion of this and the other hypotheses). The relationship between consequences of models' behaviors and subjects' imitation responses has not always been symmetrical, however. In the study reported by Bandura et al. (1963b), children displayed more imitative aggression after they had seen an aggressive model rewarded for aggression, but children who viewed an aggressive model punished for aggression did not reduce their level of imitative aggression, below that of the control group who viewed a nonaggressive model or the control group that did not view a model.

The remainder of our hypotheses concerned the effects of sex of model and sex of experimenter on the imitative responses of first-grade children. Experimenters and models were expected to produce more imitative responses in reward conditions and counterimitative responses in punishment conditions when they were the same sex as the subject. Furthermore, it was predicted that male models and experimenters would produce more imitation or counterimitation than female models and experimenters.

From a theoretical perspective, it is possible to predict that either males or females would be the more salient models and experimenters for young children. From a reinforcement viewpoint, some researchers and theorists argue that the female may be the more salient since children are exposed to females more than to males (Maccoby, 1959). In addition, mothers and teachers, usually female, reinforce children for imitating more than fathers and other males do (Rosenblith, 1961). However, the father, as a reality enforcer, may be a more salient reinforcer and model (Epstein & Liverant, 1963). He is given more prestige in this culture (Cook & Smothergill, 1973), and males are unusual in a school setting (Stevenson, Keen, & Knights, 1963). From a sex-role development perspective, by the time a child enters school, people of both sexes are positively reinforcing the child for appropriate sex-typed behaviors, perhaps learned by imitating a like-sex model.

Children undoubtedly learn at least some of their appropriate sex-typed behaviors through observational learning from like-sex models (Mischel, 1966) as well as from direct tutorage and reinforcement (Bandura, 1969). This may result in a tendency for children to imitate models of the same sex. For instance, Garrett (1971) found that pairs of opposite-sex, first-grade children who watched a videotape of a pair of opposite-sex, junior-high-school-age models engage in either appropriate or inappropriate sex-typed behaviors tended to imitate the behaviors of their like-sex models in a similar situation. Not all studies have found such effects, however (see Garrett et al., 1974).

The present study, then, was designed specifically to investigate the effects of sex of model and experimenter on children in a non-sex typed imitation situation where the subject observes the experimenter either

reward, punish, or ignore the model's responses. In this way, we wanted to explore the power of social learning theory to predict children's modeling behavior in circumstances involving variables that are least potentially relevant to sex-role development.

Hypotheses

1. Children will imitate models who are verbally rewarded by the experimenters, counterimitate models who are verbally punished, and neither imitate nor counterimitate models who are verbally ignored.

2. Males will be more salient[2] models and experimenters and will make the most salient combination of model and experimenter when vicarious reinforcement is involved.

3. The same-sex model, experimenter, and model-experimenter combinations as the subject will be the most salient when vicarious reinforcement is involved; that is, even with overall male saliency, there will be like-sex modeling within each subject's sex.

Method

The methodology in this study is modeled after that developed by Liebert et al. (1972) in a study examining the effects of vicarious consequences and race of model on imitative performance by black children.

Subjects

Subjects were drawn from first-grade classrooms made available by the Monroe County Community School Corporation in Bloomington, Indiana. One classroom of 16 boys and 12 girls was located in a predominantly rural portion of Monroe County; this classroom was used to pretest pairs of stimulus pictures for use in the actual experiment. The subject pool from which subjects were selected for the main experiment came from two other schools: four classes (72 children) from a school serving mainly an upper-middle-class urban population and two classrooms (52 children) from a school serving a blue-collar urban population. Although the heterogeneous nature of the populations tested introduced extraneous variance into the design, the authors were forced to work with available populations. Absences reduced the final pool of subjects to 115 children—62 males and 53 females. From this pool 48 children of each sex were drawn to complete the design in a manner described below. Approximately 58% of the children who actually served as experimental subjects came from the middle-class school, 42% from the other school.

[2] Saliency is defined as the degree to which the condition produces imitative responses when the model is verbally rewarded and counterimitative responses when the model is verbally punished.

Experimenters and Models

The adult experimenters and models used in the study were graduate and undergraduate students from the Indiana University School of Education. In order to decrease the possibility that nonsex differences in appearance or behavior among the models and experimenters would produce unreliable effects, two male and two female models and two male and two female experimenters were used. The experimenters and models were balanced across the design. Several training sessions were provided prior to the experiment; in these the purpose of the experiment was explained to all experimenters and models, and they practiced their roles.

Materials

Stimulus pictures. The stimulus pictures used in the experimental phase of the study were taken from the Peabody Language Development Kit, Level 1 (1965). Fifty pairs of pictures of a variety of objects (e.g., food, vehicles, animals, etc.) were selected as having no obvious differential appeal by sex. These picture pairs were pretested by a male and a female experimenter in a group, dichotomous-choice format in the first-grade classroom described above (Garrett et al., 1974).

Ten sets of picture pairs that showed equal or nearly equal preference among both boys and girls were selected for use in the main experiment. More heterogeneity in preference was found than had been expected, but the balance of preferences between pictures by boys or girls never exceeded a 4:7 ratio.

Group Toy Preference Test. The group Toy Preference Test, developed by Anastasiow (1963), was administered to all available subjects in the two schools described above. This test uses pictures of five "masculine" and five "feminine" objects and provides a measure of current sex-role preference, which was used as a blocking variable in this design: The objects had differentiated significantly between girls and boys ($p < .001$) in research performed earlier using actual toys. Details concerning the administration and scoring of this test may be found in Garrett et al. (1974).

In evaluating the test, Anastasiow's test–retest comparisons yielded a rank-order correlation of .96 and a Pearson product-moment correlation of .81 on individual scores. Comparisons with the Wright Picture Preference Test and the Sears-Wright Toy Preference Test yielded significant correlations in test results, convincing Anastasiow that the three tests were measuring the same construct.

Procedures

The Group Toy Preference Test was administered by a male and a female experimenter to intact

classrooms approximately two weeks before the main experiment began. One experimenter read the directions and paced the subjects, while the other supervised the subjects and attempted to resolve any problems. The experimenters reversed their roles in each successive classroom so that half of the time each sex served in each role. Procedures for administration were adapted from those used by Anastasiow (1963).

Scores for each subject were computed as specified above and quartiles determined for purposes of assigning subjects to the design.

The experimental phase of the study was implemented over a period of one week during the spring of 1973. The experimental sessions were held in school hallways or unused rooms. Despite less than ideal conditions, there were few external disturbances. The experimenter went to the subject's classroom, introduced himself (or herself), and brought the child back to the experimental location. On the way, the experimenter chatted with the subject, telling the subject that he was going to "help us pick some pictures." At the experimental location, the experimenter introduced the subject to the model by name and asked the subject to sit down. The model always sat directly across a table from the experimenter, with the subject sitting by the model's side.

The experimenter then said, "I am going to show you both some pictures of objects. First I am going to show ——[the model] the pictures, and he/she will tell me which ones he/she likes. Then after—— is done, it will be ——'s [the subject's] turn to tell me which picture he/she likes best. All right, ——, which of these two pictures do you like best: the hot dog or the hamburger?"

The 10 pairs of pictures were then presented to the model, one pair at a time with the pictures held side by side. The same sequence for presenting the pictures was used throughout the experiment. The pictures were also marked so that each picture was always presented on the same side of the pair. The experimenter had a record sheet for each subject, indicating the subject's assigned experimental treatment (reward, punish, or ignore the model's choices) and the randomly selected specific choice the model was to make within each picture pair.

After the experimenter asked for the model's preference within each picture pair, the model paused about two seconds and then responded by pointing to the appropriate picture and saying "This one." No other verbalizations or actions were made by the model. Depending on the subject's treatment condition, the experimenter made one of the following responses to the model's choice:

Reward the model: "That's a good one." or "That's the better one."

Punish the model: "That's not a good one." or "That's not the best one."

Ignore the model: no response was made to the model's choice.

The experimenter also marked the model's choices on the record sheet to maintain the illusion for the subject that the experimenter was keeping a record of the model's choices.

After the model had viewed all 10 picture pairs, the experimenter thanked him; the model left the experimental location. The subject was then told, "Now, it's your turn. Tell me which of these two pictures you like the best: the hot dog or the hamburger." The experimenter then proceeded as with the model, presenting the pictures and asking for and recording the subject's preferences. All approving or disapproving comments were avoided. When the subject finished, he or she was praised for doing a good job and helping the experimenter and then returned to the classroom. Total time for administering the procedures was approximately five minutes. Only minor deviations from these procedures occurred.

Design

The design of this study was a $3 \times 2 \times 2 \times 2$ (Vicarious Consequences $\times$ Sex of Experimenter $\times$ Sex of Model $\times$ Sex of Subject) full factorial design with sex-role preference quartile as a blocking variable. This design resulted in four subjects in the highest-order interaction cell.

Analysis

The data were analyzed using analysis of variance techniques with planned orthogonal comparisons. Hypothesis 1, the main effect of vicarious consequences, was tested with Scheffé multiple comparisons. Hypothesis 2, the interactions of vicarious consequences with sex of the model (two-way), with sex of the experimenter (two-way), and with sex of the model and sex of the experimenter (three-way), was tested with planned orthogonal comparisons. Hypothesis 3, the interactions of vicarious consequences and sex of the subject with sex of the model (three-way), with sex of the experimenter (three-way), and with sex of the model and sex of the experimenter (four-way), was tested with planned orthogonal comparisons.

Results

Table 1 presents the three general hypotheses, the subhypotheses derived from these for testing the effects of sex of the model, experimenter, and subject within the vicarious reward and punish conditions, and whether the subhypotheses were statistically verified. As this table indicates, Hypothesis 1 was partly verified. The main effect of vicarious reinforcement condition was significant ($F = 15.32$, $df = 2/72$, $p < .001$). Scheffé multiple comparisons revealed that the reward ($\bar{X} = 5.10$ out of a maximum of 10) and ignore ($\bar{X} = 4.47$) conditions

TABLE 1

STATISTICAL VERIFICATION OF HYPOTHESES

Hypothesis[a]	Statistical verification[b]
1. Vicarious reinforcement condition	
(*a*) Reward ($n = 32$) > ignore ($n = 32$)	No
(*b*) Ignore > punish ($n = 32$)	Yes
2. Male *M*s, *E*s, and *E*-*M* combinations most salient	
(a) In reward condition:	
(*i*) Male *M* ($n = 16$) > female *M* ($n = 16$)	No
(*ii*) Male *E* ($n = 16$) > female *E* ($n = 16$)	No
(*iii*) Male *E*-male *M* ($n = 8$) > all other *E*-*M* combinations ($n = 24$)	No
(*b*) In punish condition:	
(*i*) Male *M* ($n = 16$) < female *M* ($n = 16$)	No
(*ii*) Male *E* ($n = 16$) < female *E* ($n = 16$)	No
(*iii*) Male *E*-male *M* ($n = 8$) < all other *E*-*M* combinations $n = 24$)	No
3. Same-sex *M*, *E*, and *E*-*M* combinations as *S* most salient	
(*a*) In reward condition:	
(*i*) Male *M*-male *S* and female *M*-female *S* ($n = 16$) > all other *M*-*S* combinations ($n = 16$)	Yes
(*ii*) Male *E*-male *S* and female *E*-female *S* ($n = 16$) > all other *E*-*S* combinations ($n = 16$)	No
(*iii*) Male *E*-male *M*-male *S* and female *E*-female *M*-female *S* ($n = 8$) > all other *E*-*M*-*S* combinations ($n = 24$)	No[c]
(*b*) In punish condition:	
(*i*) Male *M*-male *S* and female *M*-female *S* ($n = 16$) < all other *M*-*S* combinations ($n = 16$)	No
(*ii*) Male *E*-male *S* and female *E*-female *S* ($n = 16$) < all other *E*-*S* combinations ($n = 16$)	No
(*iii*) Male *E*-male *M*-male *S* and female *E*-female *M*-female *S* ($n = 8$) < all other *E*-*M*-*S* combinations ($n = 24$)	No

Note. *M* = model, *E* = experimenter, *S* = subject.

[a] The number in parentheses indicates the number of subjects involved in that portion of the comparison.

[b] Hypotheses 1*a* and 1*b* were tested by means of Scheffé multiple comparisons; all other hypotheses were tested by means of planned orthogonal comparisons, using the mean square error term from the full factorial analysis of variance as the best estimate of the pooled error term for the comparisons (Winer, 1971).

[c] Although not statistically significant, the direction of the means is in accord with this subhypothesis.

were not different, but the punishment ($\bar{X} = 1.84$) and ignore conditions were ($F = 8.84, df = 1/72, p < .01$). Significantly more imitation occurred in the vicarious ignore and reward conditions than in the punish condition.

A secondary analysis using t tests compared the sample mean of the imitation scores for male and female children in the vicarious reward and punish conditions against the population mean of 5.00, the mean value of the imitation scores that would have occurred if the treatment had had no effect on the children's choices. Only the punish condition yielded means that statistically differed from the population value (male children: $\bar{X} = 1.69, t = -6.44, p < .001$; female children: $\bar{X} = 2.00$ $t = -4.60, p < .001$).

All the subhypotheses associated with Hypothesis 2 were not supported. In the reward and punish conditions, male experimenters and models were not more salient than female experimenters and models.

Hypothesis 3 received partial support in the reward condition, no support in the punish condition. The planned comparisons indicate that in the reward condition same-sex models as the subject were more salient for young children than models of the opposite sex ($F = 4.23, df = 1/72, p < .05$); there was no difference between same- and opposite-sex models in the punish condition. There was no difference between same- and opposite-sex experimenters as the subject in either the reward or punish condition.

Because of the small cell size ($n = 4$) in the interaction of vicarious reinforcement condition, sex of the experimenter, sex of the model, and sex of the subject, multiple-comparison tests among cell means would lack power. However, an examination of the means associated with this interaction may disclose some interesting trends. Table 2

TABLE 2

MEANS AND STANDARD DEVIATIONS FOR THE INTERACTION OF VICARIOUS REINFORCEMENT CONDITION AND SEX OF EXPERIMENTER (E), MODEL (M), AND SUBJECT (S) (n = 4)

Sex	Condition					
	Vicarious reward		Ignore		Vicarious punish	
	Male S	Female S	Male S	Female S	Male S	Female S
Male M-Male E						
$\bar{X}$	6.50	4.25	4.00	3.50	2.00	3.50
SD	2.38	2.63	2.45	3.00	1.41	4.73
Male M-Female E						
$\bar{X}$	6.25	4.50	5.00	3.50	2.75	1.75
SD	4.35	1.29	2.45	3.11	2.75	1.71
Female M-Male E						
$\bar{X}$	4.25	5.00	3.50	5.00	.75	.75
SD	3.40	0.82	1.73	1.83	1.50	.96
Female M-Female E						
$\bar{X}$	3.75	6.25	5.50	5.75	1.25	2.00
SD	1.50	3.30	2.08	1.50	2.50	1.63

presents these means and standard deviations. As the table indicates, the highest imitation scores for the reward condition were obtained by female children exposed to female models and experimenters and male children exposed to male models and experimenters. So the direction of the data is in accord with the corresponding subhypothesis. For male and female children in the punish condition, the greatest counterimitation occurred with a male experimenter and female model. This trend is not in the direction specified in the corresponding subhypothesis.

The effect of the different models and experimenters was explored by examining the means and standard error of the means associated with each participant in each reinforcement condition. The mean scores of each participant generally were within ± 1 standard error of the other participants. The specific individuals involved in the study as experimenters and models did not seem to make much difference in the children's behavior. However, one female was particularly ineffective; her imitation scores in all three conditions were approximately the same.

DISCUSSION

The effects of vicarious reward and punishment were asymmetrical: that is, only punishment produced a significantly different number of imitation responses from the ignore condition. It is possible that the subjects in the various treatment groups may have interpreted the experimenter's silence with respect to their choices differently. A subject who had observed the experimenter lavishly reward the model may have interpreted the experimenter's silence as direct punishment, which could have inhibited the subject's imitation responses. A subject in the punishment condition may have interpreted the experimenter's silence as an indication that his or her responses were correct, and hence, the subject's counterimitative responses would have persisted. This interpretation is consistent with the literature on the direct consequences of telling subjects that their responses are right, wrong, or not commenting (e.g., Buss, Braden, Orgel, & Buss, 1956). The general finding in these experiments was that either the combination of "right" and "wrong" or no comment and "wrong" produced more rapid learning than when the subject was told "right" when he or she was correct and no comment when he or she was wrong. One interpretation of these results is that no comment acquires positive reinforcing properties in the combination of no comment and "wrong"; and negative reinforcing properties in the combination of "right" and no comment (Buchwald, 1959).

It is possible that similar effects were operating in the present experiment where the reinforcement was vicarious rather than direct. Travers, Van Wagenen, Haygood & McCormick (1964) have documented such effects for vicarious learning situations.

This explanation seems plausible, but it is not consistent with the fact that research using a very similar methodology to the one employed here produced more symmetrical effects for vicarious reward and punishment (e.g., Liebert et al., 1972). Explication of these inconsistencies requires further theoretical analysis and experimentation. Future experimentation should attempt to control these possibilities through careful designs.

The remainder of our hypotheses received only partial support. There was no tendency for males or females to serve as more salient models and experimenters when vicarious reinforcement was involved. Sex of the model and/or the experimenter alone does not seem to be a relevant variable in this situation. However, the match between sex of the subject and sex of the experimenter and model does seem to be relevant. The reduction in overall variance caused by the failure of the children in the reward condition to engage in much imitation is, undoubtedly, one reason why the effects were not as strong as expected. Nevertheless, certain trends in the data do reveal interesting effects. In the reward condition, like-sex models and experimenter-model combinations did tend to produce more imitation, as predicted. In the punish condition, however, the combination of male experimenter-female model produced the most counterimitation in both sexes. Like-sex imitation, therefore, appears most likely in positive, nonthreatening situations. However, in somewhat threatening, negative conditions, the child reacts most strongly to the experimenter-model combination that resembles a traditional stereotype of home life: an instrumental and authoritarian father defining as incorrect certain behaviors of a somewhat passive mother. To what extent this stereotype typifies the actual family relationships of the children involved is unknown. One might expect that such a relationship would be more prevalent among lower-class rather than middle- or upper-class children, but our sample size was not sufficient to test this assertion. Future research of this type should attempt to ascertain the relationship between the subject's performance on modeling tasks and the type of family relationships which the child or others perceive as existing in the home.

Thus, this experiment, while it contains certain departures from our expectations, does demonstrate that social learning theory and experimental methodologies derived from it can provide useful information concerning the development of sex-role behaviors. Refinements of the present experiment and inclusion of other variables such as children's perception of parental dominance, views of traditional sex-role stereotypes, birth order, locus of control, etc., promise to yield meaningful information concerning sex-role development. This information, combined with other recent research developing a similar theoretical viewpoint (e.g., Cook & Smothergill, 1973), should lead to the construction of a coherent and useful model of sex-role development and to the discovery of means to alter environments, educational and otherwise, that limit individual opportunity.

REFERENCES

Anastasiow, N. The relationship of sex role patterns of first grade boys to success in school. Unpublished doctoral dissertation, Stanford University, 1963.

Bandura, A. Social learning theory of identification processes. In D. A. Goslin (Ed.), *Handbook of socialization theory and research*. Chicago: Rand McNally, 1969.

Bandura, A. *Social learning theory*. New York: General Learning Press, 1971.

Bandura, A., & Kupers, D. J. Transmission of patterns of self-reinforcement through modeling. *Journal of Abnormal and Social Psychology*, 1964, **69**, 1–9.

Bandura, A., & McDonald, F. J. Influence of social reinforcement and the behavior of models in shaping children's moral judgments. *Journal of Abnormal and Social Psychology*, 1963, **67**, 274–281.

Bandura, A., Ross, D. & Ross, S. Imitation of film-mediated aggressive models. *Journal of Abnormal and Social Psychology*, 1963, **66**, 3–11. (a)

Bandura, A., Ross, D., & Ross, S. Vicarious reinforcement and imitative learning. *Journal of Abnormal and Social Psychology*, 1963, **67**, 601–607. (b)

Buchwald, A. M. Experimental alteration in the effectiveness of verbal reinforcement combinations. *Journal of Experimental Psychology*, 1959, **57**, 351–361.

Buss, A. H., Braden, W., Orgel, A., & Buss, E. Acquisition and extinction with different verbal reinforcement combinations. *Journal of Experimental Psychology*, 1956, **52**, 288–295.

Cook, H., & Smothergill, D. W. Racial and sex determinants of imitative performance and knowledge in young children. *Journal of Educational Psychology*, 1973, **65**, 211–215.

Epstein, R., & Liverant, S. Verbal conditioning and sex-role identification in children. *Child Development*, 1963, **34**, 99–106.

Fryrear, J. L., & Thelen, M. H. Effect of sex of model and sex of observer on the imitation of affectionate behavior. *Developmental Psychology*, 1969, **1**, 298.

Garrett, C. S. The effects of modeling on the development of sex-role behaviors in children. Unpublished doctoral dissertation, Iowa State University, Ames, Iowa, 1971.

Garrett, C. S., Cunningham, D. J., & Buelow, J. O. Effects of vicarious consequences and model and experimenter sex on imitative behavior in first-grade children. Bloomington: Indiana University Institute for Child Study, 1974.

Liebert, R. M., Sobol, M. P., & Copemann, C. D. Effects of vicarious consequences and race of model upon imitative performance by black children. *Developmental Psychology*, 1972, **6**, 453–456.

Maccoby, E. E. Role-taking in childhood and its consequences for social learning. *Child Development*, 1959, **30**, 239–252.

Mischel, W. A social-learning view of sex differences in behavior. In E. E. Maccoby (Ed.), *The development of sex differences*. Stanford, Calif.: Stanford University Press, 1966.

Mischel, W. Sex-typing and socialization, In P. H. Mussen (Ed.), *Carmichael's manual of child psychology*. Vol. 2. (3rd ed.) New York: Wiley, 1970.

Peabody Language Development Kit. Level 1. Circle Pines, Minn.: American Guidance Service, 1965.

Rosenblith, J. F. Imitative color choices in kindergarten children. *Child Development*, 1961, **32**, 211–223.

Stevenson, H. W., Keen, R., & Knights, R. M. Parents and strangers as reinforcing agents for children's performance. *Journal of Abnormal and Social Psychology*, 1963, **67**, 183–186.

Travers, R. M. W., Van Wagenen, K., Haygood, D. H., & McCormick, M. Learning as a consequence of the learners' task involvement under different conditions of feedback. *Journal of Educational Psychology*, 1964, **55**, 167–173.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

(Received March 7, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 948-955

LEARNING ABILITY AND INTELLIGENCE AS COGNITIVE PREDICTORS OF ACHIEVEMENT IN FIRST-GRADE BLACK CHILDREN[1]

TREVOR E. SEWELL[2]
Temple University

ROGER A. SEVERSON
University of Wisconsin-Madison

The study sought to examine the relationship between achievement and a number of variables shown to be associated with learning ability, including IQ, to determine their predictive effectiveness. Sixty-two first-grade black children were individually tested on the Wechsler Intelligence Scale for Children and tasks from three learning assessment strategies: (*a*) learning potential strategy using the Raven's Coloured Progressive Matrices in a pretest-coaching-posttest format, (*b*) diagnostic teaching, and (*c*) paired-associate learning under three conditions designed to facilitate learning. IQ correlated moderately with achievement. In general, diagnostic teaching exceeded IQ in predictive effectiveness. Prospects for a more precise determination of the learning potential in young children were discussed.

The limitations of the intelligence test lie not only in its inability to accurately differentiate the learning potential of children, but in the value of the information for prescribing educational change. This increasing concern about the predictive and diagnostic effectiveness of the IQ test has been focused primarily on the performance of subcultural populations undoubtedly because of the conviction that its use often leads to highly questionable social and educational practices.

Several approaches to the assessment and prediction of learning ability have been advocated to circumvent the limitations of the intelligence test. Among these are the development of systematic teacher observational techniques and the use of achievement tests (Loretan, 1965; Yourman, 1964). Other areas of investigation that seem to hold promise for advancing beyond the limitations of IQ are (*a*) learning potential assessment strategy, (*b*) diagnostic teaching, and (*c*) paired-associate learning.

Focusing almost exclusively on the educable mentally retarded, Budoff and his associates (Budoff & Friedman, 1964; Budoff, Miskin, & Harrison, 1971) developed a strategy geared to assess the *learning potential* of individuals from non-middle-class backgrounds by exposing the individuals to actual learning experiences. The strategy was based on the hypothesis that by offering the relevant learning opportunity through systematic exposure to nonverbal reasoning tasks, educable mentally retarded children could be differentiated more precisely than by the verbally biased IQ tests. On the Kohs Block Design Test he identified children as "high scorer," "gainers," and "non-gainers" on the basis of their initial high scores and ability to profit from a coaching experience. In contrast to the non-gainers with equivalent IQ, the high scorers and gainers, who were shown to be associated with disadvantaged background, were found to be educationally rather than mentally retarded. In essence, learning potential status was more predictive of learning ability with

[1] The study was conducted as a partial requirement for the first author's doctoral dissertation at the University of Wisconsin, Madison. The authors are indebted to A. Kaczmarek, and to the first-grade teachers of Auer Avenue School, Milwaukee, for their kind cooperation. The valuable assistance of Robert Davidson and Frank Farley is appreciated.

[2] Requests for reprints should be sent to Trevor E. Sewell, Department of School Psychology, Temple University, Philadelphia, Pennsylvania 19122.

lower-socioeconomic children than was the traditional IQ test.

The pressing need to recognize individual differences in learning ability not tapped by intelligence has stimulated some consideration for a *diagnostic teaching* approach to assessment (Meyer & Hammill, 1969; Rosenberg, 1968; Severson 1971). The advocates of diagnostic teaching are emphatic in their objective of shifting from an examination of static cognitive abilities toward the assessment of the process learning behavior of the child. More specifically, this involves an assessment of the individual differences that significantly influences the child's learning skills and his learning style. Empirical studies to determine the predictive and diagnostic value of diagnostic teaching are still in an exploratory stage. One such study is that of Jones (1970) who investigated the relationship between learning ability and achievement. She reported a correlation of .73 which underscores the need for further examination of diagnostic teaching in comparison with other variables shown to be associated with learning ability.

Jensen (1961) initiated awareness of the importance of looking at actual learning abilities of subcultural populations primarily because of the influence of past learning on IQ performance and the pronounced differences in environmental conditions for school-related learning between different socio-economic status (SES) groups. In the efforts to separate the individual's past learning and current learning abilities as reflected in the IQ score, the method of assessing the learning ability using *paired-associate tasks* which provide direct measures of learning abilities was conceived. The wide disparity between blacks and whites, lower- and upper-socioeconomic classes in measured intelligence is not reflected in an equally discrepant learning ability when learning is measured directly by performance on paired-associate tasks (Green & Rohwer, 1971; Rohwer, Ammon, Suzuki, & Levin 1971; Rohwer, Lynch, Levin, & Suzuki, 1968; Semler & Iscoe, 1963). Intrigued by these findings, psychologists and educators began to question the relationship between laboratory-based, paired-associate learning tasks and classroom learning. Resulting research have shown a moderate relationship between paired-associate learning and school achievement (Giebink & Goodsell, 1968; Lambert, 1970; McCullers, 1965; Rohwer & Levin, 1971; Stevenson, Hale, Klein, & Miller, 1968).

In addition to having the child engaged directly in learning in order to assess his learning ability, others have asked whether the child could improve his learning ability as a function of training (Davidson, 1964; Rohwer, 1971). The appropriateness of training activities to enhance learning can be seen from the observation that the culturally disadvantaged child on nonverbal tests reveals an absence of the mediational tendency found among the middle- and upper-class children. Jensen (1966) noted that the tendency to verbally mediate is one aspect of learning ability that is susceptible to environmental influence. Essentially the same observation was made by Rohwer (1971) who pointed out that differences in paired-associate performance by different SES groups can be attributed to differences in spontaneous elaboration by children of different populations.

The effectiveness of mediators in facilitating paired-associate learning has been adequately demonstrated (Davidson, 1964; Jensen & Rohwer, 1963; Paivio & Yuille, 1967; Rohwer & Ammon, 1971). Consequently, if one asumes that the culturally disadvantaged child has had less than an optimal opportunity for the development of his verbal facilities, then conceivably, performance on paired-associate learning tasks after appropriate training in elaborative process would not only be important to school achievement as Jensen (1966) suggested but could provide invaluable educational diagnostic information.

The present study is an attempt to examine the relationship between achievement and number of variables shown to be associated with learning ability, including the individually administered intelligence test to determine their predictive effectiveness. A secondary objective is to investigate the interrelationships between the three methods of assessing learning ability (learning po-

tential assessment, paired-associate and diagnostic teaching) selected to focus on the shift from the assessment of cognitive abilities by putatively culturally biased instruments to the emphasis of learning ability assessed by direct engagement in learning.

Method

Subjects

Sixty-two black children (31 boys and 31 girls) were randomly selected from five first-grade classrooms of a public elementary school in Milwaukee, Wisconsin. Subjects were from lower-socioeconomic backgrounds determined by criteria of income, residential area, and housing type. From the original sample, two subjects were excluded because of apparent sensory handicaps and three transferred to different schools before all tasks were administered. The age range was 5 years 10 months to 7 years 5 months with a mean of 6 years 4 months.

Materials and Procedure

Learning potential assessment strategy. This approach, (Budoff & Friedman 1964) involves an initial administration (pretest) of the Raven's Coloured Progressive Matrices (CPM), Sets A, AB, and B, followed by training to achieve the following objectives: (*a*) help the child to attend to all choices presented in each problem; (*b*) encourage reflective thinking rather than impulsive choosing of an answer; and (*c*) to aid in developing an accurate analogy in the 2 × 2 format employed. The training format consisted of a number of important components and strategies. No actual test items were used in training, and examples consisted of pictures of real objects (tree, car, flag, etc.). Great effort was made to explain the problem-solving strategies, and motor involvement was insured by having the child draw the item that completes the pattern before choosing the various alternative solutions presented. Following training, a posttest was administered.

Paired-associate test. The material for the paired-associate tasks consisted of three lists of 25 noun–picture pairs. The pictures were simple line drawings of familiar objects taken from primer workbooks for first-grade children. Subjects were shown the picture pairs for five seconds and told that the two pictures "always go together." For each of three conditions a large sheet was created which included all response members of the picture pairs. After the stimulus picture was presented, the subject was required to point to the response object. This recognition paired-associate method (Davidson, 1964) is especially appropriate for this sample in that it eliminates the need for a verbal response. The rate of presentation on both study and test trials was five seconds, with a two-minute interval between trials. The number of correct responses on the test trial was used as the dependent variable.

Condition 1 involved presentation of side-by-side picture pairs on an 8½ × 11 page. Children were instructed to associate each pair so they could remember the second one when the first was presented, and each picture pair was shown once.

Condition 2 was identical to Condition 1 except that subjects were instructed to generate an elaboration strategy to enable them to remember the pair (e.g., "Make a sentence out of the paired objects").

Condition 3 was an imposed elaboration condition wherein a picture was shown depicting an interaction between both objects in addition to the side-by-side picture. The investigator then read a sentence which linked the picture pair in the interaction (e.g., "The *present* covers the *leaf*"). Both pictures were presented on the same page and instructions were provided on the first trial only.

Diagnostic teaching. The material for the examination of the changes in behavior of the child under different conditions of reinforcement was 15 words unknown to the subjects. They were presented a list of 25 words of approximately equal learning difficulty in order to select 15 which the child could not decode. The words were slightly idiomatic in order to reduce probability of familiarity, but chosen so they would be in the speaking vocabularies of most first-grade children (Severson, 1971).

Five words were selected for each of the three different teaching conditions which proceeded from *feedback* to *social praise* to *tangible reinforcement*. All words were printed in capital letters on 3 × 5 inch index cards. Each child was exposed to the words by the following study–test method:

> I am going to show you these five words and tell you what they are. Let's see how many you can learn when I show them to you. Okay?
>
> This word is ___ (say slowly). This word is ___, this is ___, this is ___, and this is ___, (now shuffle them). Let's see how many you remember.

Under the *feedback* condition, the investigator's reaction was limited to telling the child if the response was correct or incorrect. A correct response under the *social praise* condition brought an enthusiastic, positive reaction from the investigator. During the Condition 3, the child was given a piece of candy for every two correct responses he made. The words under all three conditions were administered for a total of three complete trials, or two trials with all words completely known.

Seven subtests of the Wechsler Intelligence Scale for Children (WISC) were administered according to standard procedure. The WISC and all three measures of learning ability were individ-

ually administered to all subjects by a black examiner. The first session was devoted to the WISC, the second to the diagnostic teaching, and to the pretest and coaching on the Raven's CPM. The third session was devoted to the posttest on the Raven's CPM and the paired-associate tasks. The third session was, in all cases, on the day following the second session. Testing began the last week of September and continued through the first week of January. Achievement criterion was the Word Reading and Arithmetic subtests of the Stanford Achievement Test administered the third week in April.

Results

All achievement, WISC and learning ability variables were intercorrelated by product-moment correlations. The results are summarized in Table 1.

From Table 1 it can be seen that intelligence is related to all three conditions of diagnostic teaching, paired-associate learning, and to both pre- and posttests of the Raven's CPM. Interestingly, the imposed elaboration condition of the paired-associate testing and the posttest following training on the Raven's CPM produced substantial shifts in the relationships to intelligence and diagnostic teaching. The observed correlational differences between the Raven's CPM pre- and posttests and the other ability variables are in line with expectation and support the diagnostic significance of this approach. The three conditions of diagnostic teaching are moderately related but still show considerable independence considering the degree of similarity in each task.

Although the Verbal and Performance IQs are not as highly related to achievement as Full Scale IQ, there is clear evidence to suggest that intelligence relates to achievement at a highly significant level ($p < .001$) for black children. Of particular importance in this study is the relationship between the three subtests of diagnostic teaching and achievement. It is apparent that five of the six reported correlations between diagnostic teaching and achievement were higher than the relationship found between intelligence and achievement. Although the correlations were generally higher, only one relationship was significantly higher, that of *social praise* to arithmetic compared with *intelligence* to arithmetic ($t = 2.10$, $p < .05$). The pretest and posttests differences in the correlation of the Raven's CPM to achievement is one of the most intriguing results of the study. The strength of the relationship between posttest and achievement ($r = .47$ and $.54$, $p < .001$) is derived primarily from the gain scores (post- minus pretest). The pretest scores on the Raven's CPM show a weak relationship to achievement ($r = .22$ and $.25$).

The paired-associate tasks, either initially, following instruction to elaborate, or with imposed elaboration, show very weak relationships with achievement.

Factor Analysis

All ability and achievement variables were subjected to a principal component analysis with varimax rotation and components with eigenvalues of one or greater were retained. The six factors that emerged are summarized in Table 2.

Factor 1 is characterized by high loadings on IQ, diagnostic teaching, and achievement. Factor 2 loads highest on the Raven's CPM gain scores, with achievement and verbal intelligence showing significant loadings. This suggests that the gain score is an important component of process learning and something different from the skills in diagnostic teaching that are being reflected. Factor 4 shows a high loading on all conditions of elaboration on the paired-associate tasks, but no loading of any other variable, indicating the three tasks are unique measures virtually unrelated to intelligence or learning in a school-related sense. Factor 5 is a nonverbal IQ factor showing loadings both on Performance IQ and the initial Raven's CPM scores; it is age and sex related; and it also shows a loading with word reading. Factor 6 seems to be a developmental factor, most highly related to age.

Discussion

The primary objective of the present study was to determine the predictive validity of the IQ and a number of measures derived from three strategies of assessment based on direct engagement in learning. If IQ is the best index of learning ability for black children, then that ability should be

TABLE 1

INTERCORRELATION AMONG WECHSLER INTELLIGENCE SCALE FOR CHILDREN, LEARNING ABILITY, AND ACHIEVEMENT VARIABLES

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Sex (male; female)	—															
2. Age	.08	—														
3. Verbal IQ	−.07	−.19	—													
4. Performance IQ	−.33**	−.34**	.19	—												
5. Full Scale IQ	−.23	−.36**	.80***	.70***	—											
6. Feedback	.08	.12	.28*	.32**	.38**	—										
7. Social praise	.22	.14	.46***	.39**	.53***	.79***	—									
8. Tangible reinforcement	.19	.10	.50***	.32**	.51***	.75***	.76***	—								
9. PA—No elaboration	−.10	−.12	.06	.25*	.18	.24	.22	.25*	—							
10. PA—Instruction to elaborate	−.19	−.15	.11	.26*	.24	.27*	.21	.22	.62***	—						
11. PA—Imposed elaboration	−.27*	−.09	.21	.43***	.38***	.36**	.25*	.27*	.37**	.41***	—					
12. Pretest Raven's CPM	−.18	.19	−.007	.40**	.19	.23	.19	.16	.11	.08	.41***	—				
13. Posttest Raven's CPM	−.26*	.09	.26*	.31**	.28*	.27*	.43***	.40***	.35**	.25*	.40**	.53***	—			
14. Learning potential gain	−.19	.02	.32**	.13	.22	.17	.38**	.37**	.35**	.25*	.24	.05	.87***	—		
15. Reading	−.14	−.18	.42***	.44***	.55***	.52***	.63***	.59***	.14	.25*	.12	.22	.47***	.42***	—	
16. Arithmetic	−.13	.01	.46***	.40***	.50***	.58***	.69***	.62***	.16	.16	.17	.25*	.54***	.49***	.73***	—

Note. PA = paired-associate; Raven's CPM = Raven's Coloured Progressive Matrices.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

TABLE 2
FACTOR LOADINGS FOR ALL VARIABLES (PRINCIPAL COMPONENTS WITH VARIMAX ROTATION)

Variable	Factors					
	1	2	3	4	5	6
Sex					−.60	
Age					−.44	.72
Verbal IQ	−.45	.40				−.49
Performance IQ	−.31				.80	
Feedback	.90					
Social praise	.89					
Tangible reinforcement	.86					
PA—No elaboration				.90		
PA—Instruction to elaborate				.82		
PA—Imposed elaboration			.87	.36		
Pretest Raven's CPM			.30		.45	.68
Posttest Raven's CPM		.81				.30
Learning potential gain score		.91				
Reading		.34			.36	
Arithmetic	.68	.43				
Percentage of total variance	.73	.43				
	27.9	12.7	11.7	10.9	10.3	7.9

Note. ± .30 and above factor coefficient used as criterion of acceptance. PA = paired-associate; Raven's CPM = Raven's Coloured Progressive Matrices.

reflected in a high correlation with achievement. On the other hand, if the premise that learning ability is best assessed by having the child engage directly in learning tasks, then a substantial correlation with achievement should be expected from these strategies.

The results are clearly consistent with reported predictive effectiveness of the IQ (Jensen, 1969). However, if the prediction of achievement is the primary purpose for testing, diagnostic teaching with its criterion-referenced quality seems decidedly preferable to the IQ in the light of these results. Diagnostic teaching which emphasizes a shift from static cognitive abilities to a more process-oriented assessment is as effective a predictor of achievement as the IQ. In fact, it accounted for approximately twice the variability in arithmetic as the IQ measure.

The identification of a possible alternative predictive measure has obvious implications to the discussion concerning intelligence testing of low-SES children. If the IQ is accepted as an accurate index of learning ability, inferentially there would be a discernible similarity in learning potential between lower-class and middle-class children of equivalent IQ. The available evidence indicates (Jensen, 1968) that children with low IQs from middle-class backgrounds are invariable slow learners in contrast to low-SES children of equivalent IQs who demonstrate a wide range of learning ability. Although the predictive effectiveness of the low mean IQ (87) is evident, the indications that it is not an accurate index of learning ability for blacks (Semler & Iscoe, 1963) and the indications that "the labeling of a child with a 'permanent' stratification index (IQ) is likely to affect his self-concept, his goals, his motivations and his achievement [Yourman, 1964, p. 109]" should militate against widespread use with black children.

Perhaps most intriguing in this study was the results of the learning potential strategy used in this context. The Raven's CPM has been referred to as a test which measures "g" or general intelligence in "nearly pure form" (Jensen, 1969) since it presumably minimized cultural and scholastic content. In the present study, it showed a weak relationship to achievement when used in the learning potential pretest format. However, its effectiveness in identifying the students'

ability to profit from instruction was clearly demonstrated in the posttest results following training. Whereas the initial testing on the Raven's CPM accounted for 6% of the variability in achievement, posttesting following teaching of the relevant concepts to solve the problems accounted for approximately 25%. Interestingly, the difference between the pretest and posttest (labeled "Learning Potential Gain") was equally effective as the posttest in its predictive powers. The strategy did not exceed the IQ test in predictive potential, nevertheless, the procedure does lend itself to the prospect of better differentiating those who can profit from learning experience than the IQ measure.

Optimism regarding the prospect of paired-associate tasks adding meaningfully to the prediction of academic achievement has increased by a number of studies (Lambert, 1970; McCullers, 1965; Rohwer & Levin, 1971). Despite the significant correlation found between one condition of the paired-associate (instruction to elaborate) and reading ($r = .248$, $p < .05$), the present factor analysis and correlational data do not support the findings of a substantial correlation between paired-associate learning tasks and achievement. In light of the generally low correlations, these findings offer support to Rohwer, Ammon, Suzuki, & Levin (1971) who argued the possibility "that paired associate tasks do not elicit the kinds of learning processes necessary for performance on school learning tasks [p. 13]."

Because the patterns of predictive relationships are so varied, even among tasks which measure the actual learning process, factor analysis has contributed to the determination of important common processes involved among the various tasks. The most potent factor was obviously related to achievement, and all measures which related to achievement load highly on this factor. The Factor 2 does offer support for the independent contribution of the Raven's CPM to achievement. This factor seems to be strongly related to language, and very possibly reflects a verbal mediation skill, as opposed to the rote learning skill reflected in the first factors. None of the subsequent factors seem very important when considering achievement. They add support for the complete independence, and possible insignificance, of the paired-associate tasks when considering the kinds of achievement skills important for first-grade, low-SES black children.

Although this study did not show the WISC test to be less predictive of a black sample than is commonly reported for white children, it did offer added information regarding the value of a shift toward a curriculum-based assessment. It also underscores the value of studying the child as he actually learns, as opposed to simply assessing what the child knows from past experiences. By extending these two qualities, observing the child under conditions of standardized learning, and by selecting material from the actual body of material to be learned by the child in the curriculum, some of the limitations of the IQ test should be circumvented. The present procedure does not deal with the issue of which curriculum is the most appropriate for a given subculture, but it raises vital questions concerning the practice of testing children on materials where the possibility of content irrelevance is considerable and where the historical experiences of the child have poorly equipped him to deal with standardized tests.

REFERENCES

Budoff, M., & Friedman, M. "Learning Potential" as an assessment approach to the adolescent mentally retarded. *Journal of Consulting Psychology*, 1964, **28**, 433–439.

Budoff, M., Meskin, J., & Harrison, R. Educational test of the learning-potential hypothesis. *American Journal of Mental Deficiency*, 1971, **76**, 159–169.

Davidson, R. Mediation and ability in paired-associate learning. *Journal of Educational Psychology*, 1964, **55**, 352–356.

Giebink, J., & Goodsell, L. Reading ability and associative learning for children with a visio-motor deficit. *American Educational Research Journal*, 1968, **5**, 412–419.

Green, R., & Rohwer, W. SES Differences on learning and ability tests in black children. *American Educational Research Journal*, 1971, **8**, 601–609.

Jensen, A. Learning abilities in Mexican-American and Anglo-American children. *California Journal of Educational Research*, 1961, **12**, 147–159.

Jensen, A. Verbal mediation and educational potential. *Psychology in the Schools*, 1966, **3**, 99–109.

Jensen, A. Social class, race, and genetics: Implications for education. *American Educational Research Journal*, 1968, **5**, 1–42.

Jensen, A. How much can we boost IQ and scholastic achievement. *Harvard Educational Review*, 1969, **39**, 1–123.

Jensen, A., & Rohwer, W. Verbal mediation in paired-associate and serial learning. *Journal of Verbal Learning and Verbal Behavior*, 1963, **1**, 346–352.

Jones, M. Paired-associate learning as a predictor of first grade reading achievement. Unpublished master's thesis, University of Wisconsin, 1970.

Lambert, N. Paired-associate learning, social status and tests of logical concrete behavior as univariate and multivariate predictors of first-grade reading achievement. *American Educational Research Journal*, 1970, **7**, 511–528.

Loretan, J. The decline and fall of group intelligence testing. *Teachers College Record*, 1965, **67**, 10–17.

McCullers, J. Correlates of verbal paired-associate learning in children. *Psychological Reports*, 1965, **17**, 747–752.

Meyer, P., & Hammil, D. *Methods for learning disorders*. New York: Wiley, 1969.

Paivio, A., & Yuille, J. Mediation instructions and word attributes in paired-associate learning. *Psychonomic Science*, 1967, **8**, 65–66.

Rohwer, W. Learning, race, & school success. *Review of Educational Research*, 1971, **41**, 191–210.

Rohwer, W., & Levin, J. Preferences and differences in learning proficiency. In J. Hellmuth (Ed.), *Cognitive studies*. New York: Brunner/Mazel, 1971.

Rohwer, W., & Ammon, M. Elaboration training and paired-associate learning efficiency in children. *Journal of Educational Psychology*, 1971, **62**, 376–383.

Rohwer, W., Ammon, M., Suzuki, N., & Levin, J. Population differences in learning proficiency. *Journal of Educational Psychology*, 1971, **62**, 1–14.

Rohwer, W., Lynch, S., Levin, J., & Suzuki, N. Pictorial and verbal factors in the efficiency learning of paired-associates. *Journal of Educational Psychology*, 1967, **58**, 278–284.

Rohwer, W., Lynch, S., Levin, J., & Suzuki, N. Grade level school strata and learning efficiency. *Journal of Educational Psychology*, 1968, **59**, 26–31.

Rosenberg, M. *Diagnostic teaching*. Seattle: Special Child Publications, 1968.

Semler, I., & Iscoe, I. Comparative and developmental study of the learning abilities of Negro and white children under four conditions. *Journal of Educational Psychology*, 1963, **54**, 38–54.

Severson, R. The case for the classroom management consultant. (Experimental Publication System Ms. No. 397-36) Washington, D.C.: American Psychological Association, 1971.

Stevenson, H., Hale, G., Klein, R., & Miller, L. Inter-relations and correlates in children's learning and problem solving. *Monograph of the Society for Research in Child Development*, 1968, **33**, No. 7.

Yourman, J. The case against group IQ testing. *Phi Delta Kappan*, 1964, **46**, 108–110.

(Received March 18, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 956–961

LOCUS OF CONTROL AND APTITUDE TEST SCORES AS PREDICTORS OF ACADEMIC ACHIEVEMENT[1]

WALTER R. NORD,[2] FRANCIS CONNELLY, AND GEORGE DAIGNAULT

Washington University

The experimenters studied the relationships among academic achievement in graduate school, perceived locus of control, and a relevant aptitude test. Grades in 15 individual courses and overall grade point average (GPA) were used as the criteria of academic success. Three I–E scores (total scores and ideological and personal control subscores) and the Admissions Test for Graduate Study in Business (ATGSB) scores of 50 entering MBA students were used as predictors. It was found that both the ATGSB and I–E scales accounted for significant and often complementary portions of variance in GPA and grades in individual courses. Each instrument predicted better in certain courses and the predictive power of each instrument tended to remain when influence of the other was controlled statistically. Analysis of achievement as a function of the two I–E dimensions revealed that personal control tended to be a better predictor of achievement than was ideological control. It was concluded that the locus of control might be a useful predictor of academic success, but that no simple pattern exists. Rather, course content and such factors as teacher behavior may interact with perceived locus of control to determine academic achievement.

The degree to which perceived locus of control can be used to predict academic achievement has been given increased attention in the last few years. Comprehensive reviews of the literature (Lefcourt, 1966; Rotter, 1966) reported some support for Rotter's I–E scale as a predictor of academic achievement at the precollege level, but little evidence concerning its usefulness for college or graduate school populations.

Recently several investigators have studied the relationship of perceived locus of control and performance in college. Eisenman and Platt (1968) found no relationship between I–E scores and grades in psychology classes for 131 students, mostly freshmen, at the University of Georgia. Warehime (1972) using freshman subjects at the University of Iowa, found that the I–E scale had no practical use as a predictor of first-year grades above what could be accounted for by IQ alone. Massari and Rosenblum (1972) examined the relationship between the locus of control and academic achievement in college students as measured by a multiple-choice final exam for 43 female and 90 male introductory psychologists at Temple University. In addition to Rotter's I–E scale, these researchers used a revised version of the Crandall Intellectual Achievement Responsibility (IAR) scale and Rotter's Interpersonal Trust scale. They reported that neither the I–E nor the IAR scale were significantly correlated with academic achievement for men. Moreover for women, the correlations were significant but opposite from the predicted direction. In other words, the external women got higher grades than internal women. Also, the authors found no relationship between the Scholastic Aptitude Test and the IAR or the I–E scales. Hjelle (1970), using introductory psychology students at Villanova University as subjects, reported "marginal support" for the hypothesis that internal students got better grades than

[1] The authors wish to thank Powell Niland and Linda Mentler for their help in collecting the data for this study.

[2] Requests for reprints should be sent Walter Nord, Graduate School of Business Administration, Washington University, St. Louis, Missouri 63130.

external students ($\chi^2 = 2.98$, $p < .25$). Eilersen (1972) using 116 community college students enrolled in introductory psychology found that internality was predictive of achievement and participation in unstructured (as opposed to structured) classes. Thurber (1972) failed to find support for the hypotheses that the externally oriented subjects would perform more adequately on examinations based on classroom lectures and assigned reading material and that internal subjects would show greater relative achievement on more broadly defined examinations based on information which was available but not specifically assigned or discussed in a classroom.

While most of the studies summarized so far have given little support to the usefulness of the I–E scale as a predictor of academic success in college, several more recently published studies have reported stronger relationships. Brown and Strickland (1972), using introductory psychology students as subjects, found internality was associated with high grades for males but not for females. Similar results were reported by Boor (1973) in a study published after our analysis had been completed. Boor found that with intelligence partialed out, total I–E scores and Mirels (1970) Factor II or the political locus of control subscale of the I–E were significantly related ($r = -.23$, $r = -.27$, respectively) to examination scores in an introductory psychology course for males; the same statistics calculated on data from female subjects were not significant. Finally, using college students as subjects, Gozali, Cleary, Walster, and Gozali (1973) reported that internality was associated with functional behavior on achievement tests.

While the findings are mixed, existing evidence seems to indicate that locus of control, at least in some circumstances, may be a useful predictor of academic success in college for males. However, most of the studies were based on introductory psychology students and often the criterion has been limited to either grades in introductory psychology or to overall grade point averages of college freshmen. The relationships between perceived locus of control and performance at levels beyond the freshman year and in specific courses other than psychology have not been studied.

The present study investigated the I–E scale as a predictor of academic success in a graduate business school. In addition to studying a different population of students in a different environment than previous research, this study sought to examine the I–E scale as a predictor for performance in various courses as well as to compare its predictive power with a widely used aptitude test, the Admissions Test for Graduate Study in Business (ATGSB). Specifically, the study sought to determine if a personality variable such as the locus of control could improve the prediction of overall grade point average (GPA) over and above predictions that could be made by using a standard aptitude test (ATGSB). Second, the researchers sought to examine the predictive power of the I–E scale and the ATGSB for individual courses. Finally, the study explored the ideological and personal control subscales of the I–E (Gurin, Gurin, Lao, & Beattie, 1969) as predictors of grades.

Method

Subjects

Full-time students in the introductory Organizational Behavior course, in the Graduate School of Business at Washington University served as subjects. Students whose native language was not English and students who did not complete the first year of the program were excluded from the analysis, leaving 48 male and 2 female students who produced usable data.

Procedure

Students were asked to complete the I–E scale during the first class session of Organizational Behavior in the fall of 1971. They were told that participation was voluntary, that the results of the study would not be examined until they had completed the first year, and that the individual results would be made available to them during the second year of the program. The I–E scale was scored using Rotter's (1966) scoring key. The subscale scores were obtained using the scoring system of Gurin et al. (1969). Data on the ATGSB and course performance were collected from the school files.

The data were analyzed through correlational techniques. Pearson product-moment and multiple correlations were calculated to predict individual

course grades as a function of ATGSB and the I–E total scores. In addition, partial correlation coefficients were calculated for individual course grades as a function of the I–E, holding ATGSB constant and as a function of the ATGSB, holding I–E score constant. The same set of correlations was calculated for GPA as a function of ATGSB and I–E scores. Finally, the correlations of the ideological control and personal control subscales with individual course grades, with ATGSB scores removed statistically, were calculated. In all cases one-tailed tests were used to assess the significance of the correlation coefficients.

Results

The raw, multiple and partial correlations of the ATGSB and I–E scores with GPA and grades in individual courses are shown in Table 1. The bottom line in the table shows that both ATGSB and I–E scores were significantly related to GPA ($r = .38$, $p < .05$; $r = -.32$, $p < .05$, respectively). In other words, students who had high ATGSB scores and those who tended to be internal achieved higher grade point averages than students who had low ATGSB scores and those who tend to be more external. While the ATGSB and I–E scores were slightly but not significantly correlated ($r = -.18$, $p < .05$), when I–E scores were partialed out, the ATGSB scores were still significantly correlated with GPA ($r = .34$, $p < .05$); when ATGSB scores were removed statistically, the correlation of I–E and GPA remained significant ($r = -.27$, $p < .05$). Moreover, as the significant R^2 indicates, these two measures taken together account for over 20% of the variance in GPA, as compared to 14% for the ATGSB taken alone and 10% for the I–E taken alone. Thus, while the ATGSB was a slightly better predictor of GPA than the I–E, the two taken together produced better predictions than either taken alone.

The correlations of the ATGSB and I–E scores with grades in individual courses are also presented in Table 1. The simple correlation coefficients followed a pattern similar to that of overall grade point averages. Grades in most of the courses were positively correlated with the ATGSB scores and negatively correlated with I–E scores. While the size of the correlations was not large for either of the instruments, on the average the ATGSB appeared to slightly outpredict the I–E scale. Twelve of the 15 correlations between ATGSB scores and course grades were positive; 8 were statistically significant. Fourteen of the 15 correlations between I–E scores and course

TABLE 1

Correlations of I–E, Admissions Test for Graduate Study in Business (ATGSB), and Grades

Course	R			R^2 (R)	partial r	
	n	ATGSB	I-E	ATGSB+I-E	ATGSB	I-E
Managerial accounting	50	.49*	−.39*	.33* (.57)	.46*	−.36*
Financial accounting	46	.25*	−.32*	.15* (.39)	.22	−.30*
Economics (micro)	48	.22	−.22	.08 (.28)	.19	−.19
Economics (macro)	49	.12	−.13	.03 (.17)	.10	−.11
Computer programming	49	.24*	−.12	.06 (.24)	.23	.02
Marketing	49	−.05	−.47*	.22* (.47)	−.06	−.47*
Organizational behavior	48	.14	−.18	.05 (.22)	.11	−.16
Quantitative courses						
Linear programming	50	.34*	−.29*	.17* (.41)	.31*	−.25*
Calculus	46	.35*	−.10	.13* (.36)	.34*	−.03
Probability theory	46	.40*	−.28*	.20* (.45)	.37*	−.22
Statistics	40	.21	−.01	.04 (.20)	.21	.04
Finance	46	.25*	.01	.06 (.24)	.25*	.04
Management simulation	49	−.33	−.02	.12 (.35)	−.34	−.09
Management information systems	47	.00	−.18	.03 (.17)	−.04	−.19
Production	50	.28*	−.43	.22* (.47)	.23	−.40*
Overall grade point average	50	.38*	−.32*	.21* (.46)	.34*	−.27*

* $p < .05$, one-tailed tests.

grades were negative, 6 being statistically significant. The patterns of the partial correlations were similar, with the ATGSB being significantly positively related to grades in five courses and the I–E being significantly negatively related to grades in five courses. The ATGSB outpredicted the I–E in the quantitative, the financial, and one of the accounting courses. The I–E predicted well for both accounting courses, but had its greatest power for predicting grades in marketing and production.

Seven of the multiple correlations using ATGSB and I–E scores as predictors of course grades were significant. In four courses these two predictors accounted for over 20% of the variance in grades.

In general, the data on individual course grades supported the expectations that both instruments would be predictive of grades. However, both predictors seemed to be more successful in some courses than others. Moreover, both instruments accounted for complementary portions of the variance in grades in many of the courses.

Finally, the data were examined to determine if the subscales of ideological and personal control were related to grades in different courses. Table 2 shows the correlations between grades and I–E scores with ATGSB scores partialed out. Ten of the 15 grades in individual courses were negatively related to ideological control scores; however, only the correlations for marketing ($r = -.44$, $p < .05$) and production ($r = -.38$, $p < .05$) were statistically significant. Thirteen of the 15 correlations between personal control scores and course grades were negative; 5 of these were statistically significant. Thus it appeared that there was a tendency on both ideological and personal control dimensions for internal subjects to achieve higher grades. However, the personal control dimension seemed to be a somewhat better predictor of grades than the ideological subscale. The correlations between the two subscales and GPA are shown on the bottom line of Table 2. While internal subjects of both dimensions tended to achieve higher GPAs, only the correlation between personal control and GPA was significant ($r = -.27$, $p < .05$). In short, the personal control subscale appeared to be the better predictor of academic performance.

TABLE 2

PARTIAL CORRELATIONS OF I–E SUBSCALES WITH GRADE POINT AVERAGE AND GRADES IN INDIVIDUAL COURSES CONTROLLING FOR ADMISSIONS TEST FOR GRADUATE STUDY IN BUSINESS SCORES

Course	*n*	Partial correlations with I–E subscales	
		Ideological control	Personal control
Managerial accounting	50	−.18	−.38*
Financial accounting	46	−.15	−.28*
Economics (micro)	48	.08	−.06
Economics (macro)	49	−.03	−.12
Computer programming	49	.09	−.08
Marketing	49	−.44*	−.50*
Organizational behavior	48	−.21	−.12
Quantitative courses			
Linear programming	50	−.08	−.24*
Calculus	46	.10	−.07
Probability theory	46	−.16	−.13
Statistics	40	.21	.13
Finance	46	.15	.07
Management simulation	49	−.13	−.18
Management information systems	47	−.15	−.15
Production	50	−.38*	−.33*
Overall grade point average	50	−.11	−.27*

* $p < .05$.

DISCUSSION

In interpreting the results of this study, several limitations must be considered. First, almost all of the students were males. Second, it must be remembered that the students had already entered graduate business school. Consequently, the predictive power of the ATGSB is apt to be understated in comparison to what it would be for the population of applicants to business school. Also, the students were to some degree selected on other characteristics such as undergraduate grades and recommendations. To the degree (as yet unknown) that these measures are a function of locus of control, the predictive power of the I–E scale may be understated. Finally, grades are only one criterion of academic success and are rife with problems of reliability and validity.

While these limitations mean that no firm

statements can be made about the relative predictive power of the two instruments for selection, the data did yield some potentially useful results. First, the study confirmed previous research (Connelly & Nord, 1972) as to the magnitude of the correlation of the ATGSB with grades in MBA programs. In addition, despite a small insignificant correlation between I–E scores and the ATGSB, the predictive power of both scales tended to remain reasonably strong even when the influence of the other was partialed out. In other words, the ATGSB and the I–E scale, to some degree, measured different dimensions. However, both dimensions were related to success in graduate school. While the absolute magnitudes of the correlations for both instruments were low, they were large enough to have some practical use. Moreover, since the predictive power of the instruments complemented each other, admissions decisions of professional schools could be improved by a relatively inexpensive personality measure, such as locus of control, in addition to the standard aptitude tests. Overall, these results added to the growing number of studies which have found a positive relationship between internal locus of control and personal achievement.

Perhaps the most interesting findings were the differences between the power of two instruments to predict grades in different courses. The ATGSB predicted better in the quantitative courses and to some degree in accounting. The I–E scale was more predictive in marketing; also it predicted well in accounting, production, and even one quantitative course. Quite probably a variety of factors including subject matter, organizational climate, and the behavior, personality, and grading policies of professors influence which student characteristics are associated with success. Some support for this suggestion comes from Runyon's (1973) study. Runyon found that satisfaction with participative and directive supervisory styles in industry was related to the perceived locus of control of the subordinates. Similar findings for satisfaction with teaching style might be expected.

The results of the present study may also help to explain the confiicting findings among previous studies of the relationship between locus of control and grades. For the most part, each of these studies was based on only one course (aften introductory psychology), grades assigned by only one or a very few teachers, and one criterion such as an objective test. The data from the current study suggest that the I–E is a better predictor in some courses than in others. A search for factors which influence the degree to which the I–E scale predicts academic success appears warranted.

Finally, the current data were consistent with the view that the I–E scale is not unidimensional. For the most part, personal control was a better predictor of achievement than ideological control. However, the direction of this relationship was contrary to Boor's (1973) finding that political control was predictive of grades but personal control was not. Two differences between Boor's and the current study might be responsible for these contradictory findings. First, Boor's subscales were scored using Mirels', (1970) system; the present study employed Gurin et al.'s (1969) scoring procedure. Second, Boor's criterion was limited to grades in psychology; the present study used 15 courses as criteria. However, one of these courses, organizational behavior, included a large amount of psychological material. Grades in this course, as in Boor's study, tended to be negatively correlated with ideological control ($r = .21, p < .10$). Perhaps then, the relationship of the subscales of locus of control to academic achievement is a function of course content.

In conclusion, it was not surprising that measures of aptitude and personality were predictive of academic achievement. However, the nature of this relationship was more complex than previous researchers seem to have assumed; both the measures used in this study predicted better for some courses than for others. The fact that the I–E scale was a better predictor of achievement in some courses than was the ATGSB and the complementary nature of the two instruments for predicting grades suggest that admissions officers might find it useful to evaluate the value of personality tests.

However, the results of the present study were based on data from only one school, used only grades in first-year courses as criteria, and left a large portion of the variance in grades unexplained.

REFERENCES

Boor, M. Dimensions of internal–external control and academic achievement. *Journal of Social Psychology,* 1973, **90,** 163–164.

Brown, J. C., & Strickland, B. R. Belief in internal–external control of reinforcement and participation in college activities. *Journal of Consulting and Clinical Psychology,* 1972, **38,** 148.

Connelly, F. J., & Nord, W. R. A study of the influence of undergraduate course content on the Admission Test for Graduate Study in Business as a predictor of success in graduate schools of business. Princeton, N.J.: Educational Testing Service, Brief No. 5, 1972.

Eilersen, H. J., Jr. Locus of control and achievement and participation in structured and unstructured introductory psychology courses. *Dissertation Abstracts,* 1972, **33,** 2164–A. (Abstract)

Eisenman, R., & Platt, J. J. Birth order and sex differences in academic achievement and internal–external control. *Journal of General Psychology,* 1968, **78,** 279–285.

Gozali, H., Cleary, A., Walster, G. W., & Gozali, J. Relationship between the internal–external control construct and achievement. *Journal of Educational Psychology,* 1973, **64,** 9–14.

Gurin, P., Gurin, G., Lao, R. C., & Beattie, M. Internal–external control in the motivational dynamics of Negro youth. *Journal of Social Issues,* 1969, **25,** 29–52.

Hjelle, L. A. Internal–external control as a determinant of academic achievement. *Psychological Reports,* 1970, **26,** 326.

Lefcourt, H. M. Internal versus external control of reinforcement: A review. *Psychological Bulletin,* 1966, **65,** 206–220.

Massari, D. J., & Rosenblum, D. C. Locus of control, interpersonal trust and academic achievement. *Psychological Reports,* 1972, **31,** 355–360.

Mirels, H. Dimensions of internal versus external control. *Journal of Consulting and Clinical Psychology,* 1970, **34,** 226–228.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs,* 1966, **80**(1, Whole No. 609).

Runyon, K. E. Some interactions between personality variables and management styles. *Journal of Applied Psychology,* 1973, **57,** 288–294.

Thurber, S. D. Academic achievement as a function of task structure and locus of control. *Dissertation Abstracts,* 1972, **32,** 3807–A. (Abstract)

Warehime, R. G. Generalized expectancy for locus of control and academic performance. *Psychological Reports,* 1972, **30,** 314.

(Received March 25, 1974)

Journal of Educational Psychology
1974, Vol. 66, No. 6, 962-967

AN EXPLORATION OF TWO CORRECTION PROCEDURES USED IN MASTERY LEARNING APPROACHES TO INSTRUCTION[1]

JAMES H. BLOCK[2] AND MICHAEL L. TIERNEY[3]

University of California, Santa Barbara

This study investigated the impact on college students' grades, achievement, and attitudes of the respective "correction" procedures used in Bloom- and Keller-type mastery learning strategies. Forty-four male and female students were taught European historiography using a 2 × 3 factorial, Pretest/No Pretest × Bloom-Type Correction/Keller-Type Correction/No Correction design. The findings indicate that periodic correction can improve at least the students' ability to apply the course material, but only if it is accomplished as in Bloom's strategy. The findings also suggest that pretesting can increase students' ability to apply the course material and their attitudes toward history.

All learning for mastery strategies share some basic ideas about how instruction should be designed (Block, 1971). Yet to date there has been virtually no research as to how these ideas might work singly or in concert to facilitate student learning (Block, 1974). Accordingly, the present research attempted to take but one of these ideas and to begin to explore its potential impact on student learning. The idea selected was "correction."

Each mastery learning strategy assumes that, if someone is having learning problems on one segment of instructional sequence, these problems must be cleared up before they impair future learning. Hence, each carefully monitors student learning over every segment of the instruction and periodically remediates or corrects problems as they develop (Block, 1974).

But while each mastery learning strategy possesses some provision for correction, some strategies correct very differently than others. This is especially the case with the two best known approaches to mastery learning: Keller's (1968) Personalized System of Instruction and Bloom's (1968) Learning for Mastery Strategy. In Keller's approach, correction is typically accomplished by returning the student to the original instructional materials and methods for the segment upon which he is having learning problems. In Bloom's approach, it is usually accomplished by sending the student on to supplementary instructional materials and methods that teach the problematic subject matter in ways different from the original materials and methods.

Accordingly, the present research addressed two questions about the idea of correction. First, does correction have a positive impact on students' learning? Second, does it matter how the correction is accomplished? To begin to answer these questions, the impact on college students' learning of correction as it is practiced in Keller's and Bloom's approaches was examined.

METHOD

Subjects

Forty-four upper division college students enrolled in a 10-week European historiography course were the subjects. Thirty of the students were

[1] We are indebted to Robert M. Bortnick, Edwin M. Bridges, John W. Cotton, and Chester W. Harris for their comments about this research.

[2] Requests for reprints should be sent to James H. Block, Department of Education, University of California, Santa Barbara, California 93106.

[3] Now at the University of California, Los Angeles.

history majors and 27 had taken at least one other course from the instructor. Twenty-two of the students were males and 22 were females.

Treatments

Three instructional treatments were used to teach the course. The first treatment, the *control* treatment, was the traditional lecture/discussion approach. In this treatment, subjects attended 50-minute lectures three times a week during the quarter and read six required books.

The second treatment, the *redirected study* treatment, used the traditional approach plus a Keller-type correction procedure. Once every two weeks, students in this treatment received a multiple-choice, diagnostic-progress or formative (Airasian, 1969) test on the readings and lectures for the preceding two-week period. At the next class section their tests were returned. Each student was told which items he had answered correctly, was provided with the correct answers to missed items, and was given a prescription for learning the material tested by these missed items. This prescription indicated whether the items missed were "knowledge" items or "application" items (Bloom et al., 1956). And for each missed application item, it indicated which knowledge items covered related material. The prescription then directed the student to restudy and review the original reading materials and lecture notes for the previous two-week period.

The third treatment, the *small-group study* treatment, used the traditional approach plus a Bloom-type correction procedure. Students in this treatment were given the same diagnostic-progress tests biweekly as the redirected study group, but they had their test results returned during cooperative small-group study sessions typical of Bloom-type mastery learning strategies (see Block, 1971; Kim, 1971; Lee et al., 1971).

In these sessions, groups of approximately four students would meet together one night a week for one hour. At the start of the session, each student was given his corrected diagnostic-progress test with the correct answers indicated. Then each student in turn was asked to select one of the questions that he had answered correctly and to explain in his own words why he had selected that answer. The other students were encouraged to ask questions or to justify why they felt that other possible answers might be correct. Cooperative rather than competitive discussion was stressed, and no premium was placed upon having to come to agreement with the answer that was indicated as the correct one. This procedure was followed until all of the items on the diagnostic-progress test had been discussed.

It is important to note that due to the constraints of the instructional setting, in neither correction treatment could students be formally required to use their respective correctives. Hence a host of informal techniques had to be used. The techniques ranged from constant prodding, to asking the redirected study students for periodic self-reports on attempts to correct learning problems, to taking attendance at the small-group study sessions. While more formal procedures would have clearly been desirable, these informal procedures seemed to work quite well.

Learning Measures

There were three learning criterion measures used in the experiment. One measure was the student's final letter grade in the course, based on his performance on a take-home, essay examination. Letter grades were converted to number grades according to the standard University of California procedure (A+ = 4.3; A = 4.0; A− = 3.7; B+ = 3.3; etc.). The second measure was the student's raw score on a 50-item multiple-choice achievement test. This test was composed of 60% knowledge items and 40% application items (Bloom et al., 1956) and was constructed from a table of instructional objectives developed jointly by the second author and the course instructor. The third measure was a 10-item Likert-type attitude scale adapted from the Attitude Toward Mathematics subscale reported in Husén (1967).

Validity data neither were available nor could they be gathered on the grade or the attitude measures. However, the content validity of the achievement test was established by having the course instructor review each item and by dropping and revising those items which he did not feel tapped his instructional objectives.

Reliability data were also unavailable on the grade measure. The Kuder-Richardson (1937) Formula 20 reliability estimate for the achievement test was .78 based upon the pretest scores and .69 based upon the posttest scores. The two-day delay, test–retest reliability of the attitude measure was estimated to be .98.

Design

One third of the subjects were randomly assigned to each of the three instructional treatments. One half of the subjects in each treatment were then pretested with the achievement measure and the attitude scale so as to establish some of each instrument's psychometric properties before using them further. The remaining subjects were not pretested. Accordingly, the resultant design for this study was a 2 × 3 factorial design: Pretest/No Pretest × No Correction/Keller - Type Correction/Bloom - Type Correction.

Procedure

On Day 1 of the study, the second author addressed the students and sought their consent to participate in the experiment. He then randomly assigned the students to the three experimental treatments. The students in each group were told that they would be treated differently, but that how they were treated would not affect their chances of earning a high grade. Then one half of the students

TABLE 1

CELL MEANS AND STANDARD DEVIATIONS FOR STUDENT ACHIEVEMENT, GRADES, AND ATTITUDE TOWARD HISTORY

Group	Criterion measures								
	Achievement treatment			Grades treatment			Attitude toward history treatment		
	Small-group study	Redirected study	Control	Small-group study	Redirected study	Control	Small-group study	Redirected study	Control
Pretesting									
$\bar{X}$	30.67	25.00	28.88	32.55	31.62	30.88	13.50	12.88	14.88
SD	3.39	6.46	4.49	3.20	3.02	5.59	1.87	4.94	1.55
	($n = 6$)	($n = 8$)	($n = 8$)						
No pretesting									
$\bar{X}$	27.57	24.33	24.83	28.50	29.33	33.33	11.71	11.00	10.83
SD	4.28	4.21	6.55	3.68	6.60	3.14	4.82	3.90	2.99
	($n = 7$)	($n = 9$)	($n = 6$)						

in each treatment were randomly selected and pretested with the achievement test and the affective scale.

From Day 1 on, the instructor taught the course as he had always taught it. Biweekly, he allowed the second author to administer the appropriate diagnostic-progress test to the redirected study and the small-group study students. Control group students were allowed to leave class early on these occasions. At the end of the ninth week, all students were administered the multiple-choice achievement test and the attitude scale without any prior notice so as to minimize the effects of last-minute cramming. They then submitted their essay, take-home examinations at the end of the 10th week. These exams were graded by the instructor, and final grades were assigned by the end of the 11th week.

Data Analysis

Originally, 10 students were assigned to each of the six cells in the experimental design. Unfortunately a number of students subsequently dropped the course. Hence complete data were obtained from only 44 subjects. A 2×3 Pearson χ^2 test of association suggested that the subject mortality was essentially random (Winer, 1971). So these data were then analyzed as if they had come from a 2×3 factorial design with unequal cell sizes.

Fixed effects, nonorthogonal, univariate analysis of variance procedures (Bock, in press) were used in the data analysis. Each analysis of variance was performed in two ways since in a nonorthogonal, factorial analysis of variance it matters whether one tests for one main effect before another. In the first way, the Pretest $\times$ Correction interaction was tested, then the pretest main effect, and finally the correction main effect. In the second way, the Pretest $\times$ Correction interaction was tested, then the correction main effect, and finally the pretest main effect. When a significant main effect for correction was found on a particular learning criterion measure in either analysis of variance, the least-square estimated effects of each correction treatment relative to the control treatment were compared.

RESULTS

Table 1 reports the within-cell means and standard deviation for the scores on the achievement, grade, and attitude measures. The analyses of variance based on these descriptive statistics indicated that the pretesting had a significant effect on student attitude toward history ($F = 5.18, 5.49$, $df = 1/38, p < .05$).[4] The analyses of variance also suggested that the correction had an effect on students' total achievement test scores that approached significance ($F = 2.84, 2.82$, $df = 2/38, p < .07$). But inspection of the least-square estimated effects of each correction treatment suggest that perhaps one approach to correction had had a very different effect on student achievement than the other. The small-group study students achieved an estimated 2.26 more items correct than the control students, while the redirected study students achieved 2.19 fewer items correct than the control students.

Hence to explore in greater detail the

[4] Here and hereafter we report the two *F*s obtained by calculating each analysis of variance in two ways. The first *F* corresponds to the *F* calculated in testing the pretest main effect before the correction main effect. The second *F* corresponds to the *F* calculated in testing the correction main effect before the pretest main effect.

different effects that the two correction treatments had seemed to have on student achievement, it was decided to split each student's total achievement test score into his subscores on the knowledge items and on the application items. Two checks were then made to determine whether this split made sense, that is, whether the application and knowledge items really tapped different aspects of student achievement.

In the first of these checks, a Pearson product-moment correlation was calculated between the students' subscores on the application items and their subscores on the knowledge items. If the application and knowledge items tapped different dimensions of student achievement, then the linear correlation between scores on one set of items and scores on the other set should have been low. The correlation obtained was moderately low ($r = .36$) and only marginally different from zero ($p < .05$).

In the second check, Guttman scalogram analysis (Edwards, 1957) was performed on the knowledge and application items to determine whether knowledge of the course material was necessary, but not sufficient, for its application, as the work of Bloom et al. (1956) would suggest. Median subscores on the knowledge and on application items were calculated. Each student was then given a score of 1 or 0 depending upon whether his application subscore was above or below the median application subscore, and another score of 1 or 0 depending upon whether his knowledge subscore was above or below the median knowledge subscore. Next the Guttman coefficient of reproducibility was calculated across students under the assumption that each student who scored 0 on the knowledge items and 1 on the application items disconfirmed the hypothesis that knowledge of the course material was necessary, but not sufficient, for its application.

The subscores of only seven students exhibited the 0, 1 pattern; the coefficient of reproducibility was .84. This coefficient is slightly lower than the minimum level of .90 that Guttman has set out as being acceptable. But as Oppenheim (1966) has noted, .90 may be too stringent a criterion of reproducibility. Accordingly, the coefficient of .84 may reasonably be taken to indicate that the subscores on the knowledge and application items tended to form at least a weak Guttman scale.

Having established that it did make some sense to parse each student's total achievement test score into knowledge and application subscores, 2 × 3 nonorthogonal analyses of variance were then performed on each set of subscores to examine the effects of the pretesting and the correction. Table 2 reports the cell means and standard deviations of student subscores upon which these analyses of variance were based.

The analyses of variance indicated there was neither a Pretest × Correction interaction effect for subscores on the knowledge items nor for subscores on the application

TABLE 2

CELL MEANS AND STANDARD DEVIATIONS FOR KNOWLEDGE AND APPLICATION SUBSCORES

Group	Criterion measures					
	Knowledge treatment group			Application treatment group		
	Small-group study	Redirected study	Control	Small-group study	Redirected study	Control
Pretesting						
$\bar{X}$	16.17	14.50	16.25	14.50	10.50	12.62
SD	2.79	4.14	2.82	2.43	2.78	2.39
	($n = 6$)	($n = 8$)	($n = 8$)			
No pretesting						
$\bar{X}$	15.00	14.56	14.83	12.57	9.78	10.00
SD	3.70	3.71	4.58	2.15	1.79	2.53
	($n = 7$)	($n = 9$)	($n = 6$)			

items. But there were significant main effects for both the pretesting ($F = 5.59, 5.36$, $df = 1/38, p < .05$) and the correction ($F = 7.50, 7.61$, $df = 2/38, p < .01$) on the application items. Examination of the least-square estimated effects of each correction treatment indicated that only students who had received the small-group study treatment answered significantly more application items correctly than the students who had received no correction at all ($t = 2.47$, $p < .05$). The small-group study students had answered correctly an estimated 2.22 more application items than the control students, while the redirected study students had answered correctly an estimated 1.17 fewer application items than the control group students.

In sum, then, if learning was indexed in terms of the student's knowledge of the material taught, their final course grades, and their attitudes toward history, then students who received correction periodically throughout the course did not learn more than students who received no correction. This was true regardless of whether correction was accomplished by returning the student to the original instruction (redirected study) or by sending him to different instruction (small-group study). But if learning was indexed in terms of the students' ability to apply the material taught, then students who received correction did learn more than students who received no correction provided that correction was accomplished by sending the student to different instruction (i.e., small-group study).

Discussion

Two sets of findings have emerged from this research. One concerns the effects of correction on college students' learning. The other concerns the effects of pretesting. We will not discuss this latter set of findings here for such a discussion would be peripheral to our purpose. Clearly, the effects of pretesting on learning is a question that bears closer attention in future mastery learning research.

How then can the correction findings be explained? The most obvious explanation is that correction can have a positive impact on certain dimensions of student learning, but only if it is accomplished in a particular way. In particular, this study suggested that application of the course material could be significantly improved if students used a correction procedure which exposed them to supplementary instructional methods and materials rather than a one which required them to review and practice the original methods and materials.

It is hoped that future research would test this explanation using a wider variety of correction procedures and a number of dimensions of student learning. Such research might eventually make it possible to select correction procedures which, in combination with ordinary group-based instruction, are most powerful for promoting certain desired dimensions of student learning. In fact, it might even be combined with the work of Cronbach and Snow (in press) on aptitude-instructional treatment interactions to make it possible to select correction procedures which are most appropriate for a given set of learning outcomes as well as a given set of learners.

The major problem with such research is control for whether or not students use their correctives. If some treatment groups use their correctives, while others do not, then differences between treatment groups may be due to differential usage of correction procedures rather than to the type of correction used. In this study, we were forced to use informal procedures so as to ensure that the students in both groups were using their correctives. Clearly, more formal procedures, consonant with the classroom situation, must be developed. We are currently exploring several such procedures for use in a replication of this study.

REFERENCES

Airasian, P. W. Formative evaluation instruments: A construction and validation of tests to evaluate learning over short time periods. Unpublished doctoral dissertation, University of Chicago, 1969.

Block, J. H. (Ed.) *Mastery learning: Theory and practice*. New York: Holt, Rinehart & Winston, 1971.

Block, J. H. Teachers, teaching and mastery learning, *Today's Education*, 1973, **63**, No. 7, 30–36.

Block, J. H. A description and comparison of Bloom's learning for mastery strategy and Keller's personalized system of instruction, in J. H. Block (Ed.), *Schools, society and mastery learning*. New York: Holt, Rinehart & Winston, 1974.

Bloom, B. S. Learning for Mastery, *Evaluation Comment*, 1968, **1**, Whole No. 2.

Bloom, B. S., et al. (Eds.) *The taxonomy of educational objectives: Handbook 1: The cognitive domain*. New York: McKay, 1956.

Bock, R. D. *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill, in press.

Cronbach, L. J., & Snow, R. E. *Aptitudes and instructional methods*. New York: Appleton-Century-Crofts, in press.

Edwards, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.

Husén, T. (Ed.). *International study of achievement in mathematics*. Volume 1. New York: Wiley, 1967.

Keller, F. S. Goodbye, teacher . . . , *Journal of Applied Behavioral Analysis*, 1968, **1**, 79–89.

Kim, H. Mastery learning in the Korean middle schools, *UNESCO Regional Office for Education in Asia*, 1971, **6**, no. 1, Sec. 1, 55–66.

Kuder, G. F., & Richardson, H. W. The theory of the estimation of test reliability, *Psychometrika*, 1937, **2**, 151–160.

Lee, Y. D., et al. *Interaction improvement studies on the mastery learning project*. (Final report on mastery learning program, April–November 1971). Seoul, Korea: Educational Research Center, Seoul National University, 1971.

Oppenheim, A. N. *Questionnaire design and attitude measurement*. New York: Basic Books, 1966.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

(Received April 1, 1974)

Page 1

U. S. POSTAL SERVICE
STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
(Act of August 12, 1970: Section 3685, Title 39, United States Code)

SEE INSTRUCTIONS ON PAGE 2 (REVERSE)

1. TITLE OF PUBLICATION: Journal of Educational Psychology
2. DATE OF FILING: 9-29-74
3. FREQUENCY OF ISSUE: Bi-Monthly
4. LOCATION OF KNOWN OFFICE OF PUBLICATION *(Street, city, county, state, ZIP code) (Not printers)*: 1200 17th Street, N.W., Washington, D. C. 20036
5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS *(Not printers)*: 1200 17th Street, N.W., Washington, D. C. 20036
6. NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR

PUBLISHER *(Name and address)*: American Psychological Association, 1200 17th Street, N.W., Washington, D.C. 20036

EDITOR *(Name and address)*: Joanna Williams, Box 51 Teachers College, Columbia Univ, New York, N.Y. 10027

MANAGER EDITOR *(Name and address)*: Harold P. Van Cott, 1200 17th Street, N.W., Washington, D. C. 20036

7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)*

NAME	ADDRESS
American Psychological Association	1200 17th Street, N.W., Washington, D.C. 20036

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*

NAME	ADDRESS
NONE	

9. FOR OPTIONAL COMPLETION BY PUBLISHERS MAILING AT THE REGULAR RATES (Section 132.121, Postal Service Manual)

39 U. S. C. 3626 provides in pertinent part: "No person who would have been entitled to mail matter under former section 4359 of this title shall mail such matter at the rates provided under this subsection unless he files annually with the Postal Service a written request for permission to mail matter at such rates."

In accordance with the provisions of this statute, I hereby request permission to mail the publication named in item 1 at the reduced postage rates presently authorized by 39 U. S. C. 3626.

(Signature and title of editor, publisher, business manager, or owner)
Lynn D. Lyle Manager, Publications Business Services

10. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES (Section 132.122, Postal Manual) *(Check one)*

The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes: ☒ Have not changed during preceding 12 months ☐ Have changed during preceding 12 months *(If changed, publisher must submit explanation of change with this statement.)*

11. EXTENT AND NATURE OF CIRCULATION	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	ACTUAL NUMBER OF COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE
A. TOTAL NO. COPIES PRINTED *(Net Press Run)*	8,644	7,719
B. PAID CIRCULATION 1. SALES THROUGH DEALERS AND CARRIERS, STREET VENDORS AND COUNTER SALES	NONE	NONE
2. MAIL SUBSCRIPTIONS	7,534	6,531
C. TOTAL PAID CIRCULATION	7,534	6,531
D. FREE DISTRIBUTION BY MAIL, CARRIER OR OTHER MEANS 1. SAMPLES, COMPLIMENTARY, AND OTHER FREE COPIES	136	140
2. COPIES DISTRIBUTED TO NEWS AGENTS, BUT NOT SOLD	NONE	NONE
E. TOTAL DISTRIBUTION *(Sum of C and D)*	7,670	6,671
F. OFFICE USE, LEFT-OVER, UNACCOUNTED, SPOILED AFTER PRINTING	974	1,048
G. TOTAL *(Sum of E & F—should equal net press run shown in A)*	8,644	7,719

I certify that the statements made by me above are correct and complete. *(Signature of editor, publisher, business manager, or owner)* ▶ Lynn D. Lyle

PS Form 3526 July 1971

Journal of Educational Psychology
1974, Vol. 66, No. 6, 968-973

LOCUS OF CONTROL, TEST ANXIETY, AND STUDENT PERFORMANCE IN A PERSONALIZED INSTRUCTION COURSE[1]

GEORGE J. ALLEN,[2] LAURA GIAT, AND ROGER J. CHERNEY

University of Connecticut

The experimenters studied the effects of locus of control and trait and state test anxiety on final grades, short answer and essay examination performances, and other academic outcomes obtained from 51 females and 37 males enrolled in a Keller-type "personalized instruction" class. The course format emphasized student control over the rate at which participants mastered self-selected areas of instructional material through proctored oral examinations. Locus of control and trait anxiety were assessed during the first class, while state anxiety was measured immediately before every oral examination. Results indicate that students possessing an external locus of control contracted for and ultimately earned lower grades, began working more slowly, reported more state anxiety during oral assessments, and performed more poorly on a written final examination than their more internally oriented peers. Trait test anxiety was not reliably related to the academic outcome measures. However, a steady and significant reduction in self-reported state anxiety during oral examinations throughout the semester was found. Suggestions for "tailoring" course formats to specific student characteristics so as to improve academic performance were discussed.

A number of investigators have proposed operantly derived "personalized" instruction courses as alternatives to the more traditional lecture-examination format of university education (Born, Gledhill, & Davis, 1972; Keller, 1966). Most versions of personalized courses contain common elements which include (*a*) a priori specification of criteria for acceptable student performance on small segments of course material, (*b*) assessments which are nonpunitive and remediable, and (*c*) provision for individual variation in work rate (Johnston & Pennypacker, 1971). The assumption that students show enthusiasm for such courses and demonstrate increased learning as a result of participation has been empirically supported (McMichael & Corey, 1969; Sheppard & McDermot, 1970).

Most researchers have used an "engineering approach" in investigating the usefulness of personalized formats; that is, effectiveness of contingency manipulations is determined by repeatedly measuring behavioral changes of entire classes. Individual performance variability due to stable preexisting personality differences is treated as error variance.

The purpose of this investigation was to examine the behaviors of students differing on locus of control (Rotter, 1966) and test anxiety (Sarason, 1957) in a personalized course which was specifically designed to maximize students' control over their academic performance. Since perceived internal control is associated with superior use of personally relevant information (Phares, 1968), it was predicted that internally oriented students would (*a*) begin fulfilling course requirements more quickly, (*b*) earn

[1] Data analysis was facilitated by National Science Foundation Grant GJ-9 to the University of Connecticut Computer Center.

[2] Requests for reprints should be sent to George J. Allen, Department of Psychology, U-30, University of Connecticut, Storrs, Connecticut 06268.

higher course grades, and (*c*) predict their course performance more accurately than students with an external locus of control. As anxiety apparently hinders performance on complex cognitive tasks (Spielberger, 1966), highly anxious students would be expected to be more likely to (*a*) inaccurately predict their final course grades, (*b*) avoid taking test assessments for longer periods of time, and (*c*) achieve lower grades than less anxious individuals. In addition, highly anxious subjects may well manifest a greater frequency of task-irrelevant behaviors during oral examinations, thus interfering with concentration on the test questions themselves. The effect of prior academic performance on functioning in the course was also examined.

Method

Subjects

Of 101 students initially enrolled in an abnormal psychology course taught by the first author, 90 completed all requirements and psychometric test data were collected from 51 females and 37 males. These 88 subjects had obtained a mean previous semester grade point average (GPA) of 3.23, with a majority of the students being classified as seniors (54%) or juniors (40%).

Instructional Format

The course was designed to provide students with an opportunity to demonstrate knowledge of specific content areas at individually convenient periods throughout the semester. Knowledge of material presented during optional lectures was not assessed for purposes of fulfilling contracts for a grade. During the first class, each student received a manual specifying course procedures. After reading the manual, students returned signed contracts which specified the grade each planned to earn.

Students also received a 77-page booklet consisting of 19 self-contained programmed instructional packages (PIPs). Each PIP contained references to recent journal articles in specified topical areas (e.g., mental retardation, desensitization, childhood autism). The articles were designated as either required or suggested reading. Three question sets were also included: the first tapped factual aspects of each required reading; the second set required integration of specific experimental findings into general conclusions; the third set assessed knowledge of both factual and implicative aspects of the suggested readings. Each PIP covered approximately 92 pages of required and 81 pages of suggested readings.

After reading material for a particular PIP, the student contacted 1 of 11 undergraduate proctors (all of whom had earned an A in a previously taught course) to take an oral examination. During this assessment, the student answered a proctor-selected sample of at least four factual and all integrative questions. An acceptable performance on this portion of the oral earned the student 4 points. Answering questions pertaining to suggested readings earned an additional 2 points. Following the oral, the proctor recorded the duration of the student's speaking time and rated his performance on a scale from 1 (poor) to 5 (excellent). The student completed an identical but independent rating of his own proficiency and estimated the time he spent reading articles and reviewing for the oral.

To reinforce rapid completion of assessments, 2 bonus points were awarded for successfully passing an oral during the first five weeks of the semester, and 1 bonus point was added for each PIP completed during the second five weeks. The conversion of points to grades was as follows: A = 30; B = 22–29; C = 15–21; D = 8–14; and F = 7. Thus, students could earn an equivalent number of points by taking one less oral if they completed assessments early in the semester. An average of 2.28 orals were completed by the class (range = 0–7) during the semester.

Personality and Predictor Variables

Six potential predictors of academic success were employed. During the first class, each student completed measures of locus of control (Rotter, 1966), trait test anxiety(Sarason, 1957), and a state anxiety differential (Husek & Alexander, 1963). The latter measure was also administered by the proctors immediately before a student took an oral, as it is sensitive to anxiety fluctuations as a result of examination stress (Allen, 1970). Previous semester GPAs and Verbal Scholastic Aptitude Test (SAT) scores of all class members were also obtained.

Academic Outcome Measures

The grade originally contracted for was used as an index of expectancy for success (A = 4, B = 3, etc.) and was compared with the final grade which was actually earned. Another performance criterion involved determining the latency between the first day of class and the day each student passed his first oral. Latency was computed on the basis of a 6-day week, as no orals were given on Sundays.

During the final class, a "knowledge of abnormal psychology" test was administered. The examination consisted of 45 multiple-choice questions drawn from instructor manuals of three abnormal psychology texts by three professors who taught sections of this course the previous semester. Four short essay questions worth 55 points were also provided by the three instructors. Each essay contained two or three sections which were graded according to strictly defined criteria on a 5-point scale by two independent raters. Interscorer relia-

bility of total essay scores was quite high ($r = .96$), and scores of both raters were averaged to yield data for further analysis. This test was designed to assess general knowledge of important concepts commonly taught in abnormal psychology courses, rather than retention of the greatly diversified material contained in the PIPs. Students were informed that their scores on this test would not affect their grade, but were urged to do their best.

Student and proctor ratings of oral performance, number of minutes of student talking time during the assessment, and the number of hours the student reported studying for the oral were also used to assess outcome.

Results

Since the students completed a variable number of oral exams, all data collected during these occasions were averaged for each subject. This procedure also minimized intraindividual variability as a result of contact with different proctors and subject matter sources at various times throughout the semester.

Locus of Control and Academic Performance

Of 14 correlations computed between locus of control and the other measures, 6 were significant at the .01 level. These data indicated that an external orientation was positively associated with state anxiety during the first class ($r = .31$) and during oral examinations (.35), as well as negatively related to performance on the essay portion of the written examination (−.49). Students having an external perception also tended to begin completing orals later in the semester (.43), initially contracted for lower grades (−.32), and ultimately obtained poorer marks (−.46).

Second-order partial correlations between locus of control and several outcome variables were computed in order to partition out the influence of intellectual ability as measured by GPA and SAT. The simultaneous removal of the effects exerted by these variables did not significantly alter the magnitude of the previously reported relationships between externality and either course grade ($r_p = -.45$) or oral examination anxiety ($r_p = .35$). Partialing out the effect of latency attenuated the relationship between locus of control and final grade ($r_p = .31$), although the association remained significant.

Locus of control scores were divided at the median and the means for the resultant groups of "internals" and "externals" on all other variables were computed. Table 1 presents these data, as well as means for median-splits on trait test anxiety and previous semester grades. Results of multiple t tests, which are also contained in Table 1,

TABLE 1

Personality Scale and Academic Performance Means from Median-Splits of Locus of Control, Test Anxiety, and Previous Semester Grades

Measure	Internal (≤12)	External (≥13)	Low anxiety (≤7)	High anxiety (≥8)	Low GPA (≤3.2)	High GPA (≥3.3)
Locus of control	—	—	11.28	12.64	13.00	11.24
Test anxiety	7.75	8.88	—	—	7.70	8.76
State anxiety in first class	55.81	62.36*	54.36	62.35**	56.16	60.92
Averaged state anxiety in orals	64.09	71.25**	63.40	71.84**	67.73	67.70
Grade point average (GPA)	3.21	3.24	3.17	3.26	—	—
Scholastic Aptitude Test verbal	566.00	549.68	581.31	537.37*	533.17	577.51*
Contracted grade	3.49	3.18	3.51	3.19	3.15	3.55*
Final course grade	3.25	2.24*	2.78	2.73	2.05	3.26**
Latency	29.33	40.64*	34.25	34.93	47.48	24.26**
Short answer test	24.58	21.90*	23.52	22.89	21.77	24.19*
Averaged essay	18.40	12.95**	15.76	15.31	13.86	16.71
Averaged speaking time in oral	26.88	26.66	24.50	28.69*	27.01	26.96
Averaged student perfromance rating	3.39	3.44	3.47	3.37	3.19	3.65**
Averaged proctor performance rating	3.60	3.54	3.68	3.49	3.13	3.92**
Averaged reported study time	10.00	10.13	9.63	10.43	9.80	10.46

* $p < .05$.
** $p < .01$.

substantiate the correlational findings. Internals began completing orals 10 days earlier than externals and averaged 8 points higher on the total evaluation test score, with differences on both short answer and essay questions being significant.

No reliable differences between internals and externals were found on trait test anxiety, previous GPA, or Scholastic Aptitude Test (SAT). In addition, even though internals performed more effectively on the academic outcome measures, they did not spend more time studying for orals, receive better proctor ratings, or take less time to finish their orals. These data generally support the first two hypotheses.

Discrepancies between contracted and final grades were found for 12 internals (of which 6 overestimated their final grade) and 18 externals (including 17 overestimates). The mean *absolute* discrepancy for all internals was .19 as compared with .84 for externals ($t = 2.42, p < .02$), thus supporting the third hypothesis. A 2 × 2 chi-square comparing direction of discrepancy and locus of control yielded $\chi^2 = 7.95$, $df = 1, p < .01$, indicating that while, as a whole, those subjects whose contracted and final grades were discrepant tended to overestimate their final grades, external subjects reliably overestimated their final grades, while internal subjects equally overestimated and underestimated their final grades.

Test Anxiety and Academic Performance

Correlations between trait test anxiety and the 14 remaining variables yielded three reliable relationships, indicating that more generally anxious students reported higher levels of state anxiety both in the first class ($r = .48$) and during oral examinations (.43) and obtained lower verbal SAT scores (−.33) than their less anxious peers. Statistical analysis of the mean scores presented in Table 1 after grouping students by a median-split provided one additional reliable difference, with less anxious subjects spending less time in the oral examination situation. While this finding may be construed as weak evidence for the fourth hypothesis derived from test anxiety theory, the absence of an anxiety influence on any other academic outcome measure suggests rejection of all four predictions made for the anxiety variable.

The poor predictive usefulness of the Test Anxiety Scale may have been partially due to the course contingencies which were designed to minimize anxiety. Analysis of Anxiety Differential scores collected before the orals indicated that the entire class reported progressively less state anxiety throughout the semester. Mean Anxiety Differential scores decreased linearly from 71.51 on the first oral to 62.45 on the fifth. A t test of anxiety scores for the first and final oral by each student indicated that subsequent assessments were perceived as less stressful ($t = 2.45, p < .05$). The magnitude of state anxiety decrements was unrelated to initial level of trait anxiety.

Supplemental Analyses

The best single predictor of final grade was latency ($r = -.63$), indicating that the sooner a student began completing course requirements, the higher final grade he obtained. A stepwise multiple regression of the six predictor variables was used to predict final grade. Locus of control accounted for the greatest amount of variance, and GPA was the only other variable which significantly augmented predictive accuracy by raising R from .46 to .57. These same two variables were the only significant contributors to prediction of latency in a second regression analysis. Previous semester grades were reliable predictors of short answer and essay scores and both student and proctor ratings of oral proficiency.

These results point toward a similar pattern of performance by students with an internal locus of control and high grade point averages. It should be noted, however, that these two dimensions were relatively independent of each other ($r = -.10$) and that both locus of control ($r = .15$) and prior academic performance ($r = .06$) were not associated with trait test anxiety.

Discussion

The results of this study represent a replication of several previously reported findings and offer clarification of several current

ambiguities concerning student performance in personalized instruction courses. The strong negative relationship between latency and final grade is in agreement with Sheppard and MacDermot (1970), who reported that students who earned higher grades in an individualized course completed assessments at a faster rate than students who received lower grades.

The differential efficacy of personalized formats for teaching various "types" of students has only been tangentially investigated. Born et al. (1972) concluded that individualized instruction is most beneficial to students with histories of poor to mediocre academic performance. Learning outcomes in the present study, however, favored students with above average academic records, as these students began working earlier in the semester, were rated as providing more proficient performances on both oral examinations and on a written test, and earned better grades in the course.

It is interesting to note that although internal and external subjects were equivalent on prior scholastic aptitude and achievement measures, the former group produced reliably better performances on the academic outcome measures.

The usefulness of the locus of control measure in predicting academic outcome was consistent with the centrality of self-controlled student planning to the course. Phares's (1968) finding that internally oriented subjects manifest superior use of personally relevant information was extended to an academic setting by the present study. Internals made very effective use of course contingencies, particularly those which rewarded early completion of oral exams. During the first five weeks of the semester, 81% of the internals passed their first oral examination (thus earning 2 bonus points), whereas only 49% of the external group achieved this incentive. The more accurate prediction of their final grades by internal subjects also substantiates Phares's conclusion and is in agreement with findings reported by Wolfe (1972).

The consistent failure to find reliable relationships between locus of control and overall indices of GPA may be due to the widespread use of externally imposed rather than self-imposed control contingencies in traditional academic instruction (Aiken, 1963). For example, internal subjects achieved an average final course grade of 3.25, which was equivalent to their mean prior semester GPA of 3.21, while external subjects manifested a course grade decrement of one full grade ($M = 2.24$) in comparison with their previous GPA of 3.24. These data imply that internal subjects maintain a constant level of performance regardless of course format. The achievement of external subjects, however, declines sharply in a course which stresses student control of contingencies.

The lack of relationships between test anxiety and course performance is explicable in terms of the design of the course, which led to a progressive reduction of state anxiety throughout the semester. Analysis indicated that the decline in state test anxiety did not vary as a function of past or present academic performance but rather was due to the effectiveness of the course contingencies. While it may be speculated that personalized formats do alleviate anxiety to a greater extent than traditional examination courses, this question awaits further investigation.

Perhaps the data also imply that methods of education should be tailored to the characteristics of the individual because not all forms of teaching are equally effective with all students. If internal subjects can progress more rapidly and accrue more benefits from a semiautonomous, self-paced education, educators might give serious consideration to providing this option. It is necessary that not only intellectual inputs be the focus of education, but that individually appropriate motivational orientations be considered as well.

REFERENCES

Aiken, L. R., Jr. The grading behavior of a college faculty. *Educational and Psychological Measurement*, 1963, **23**, 319–322.

Allen, G. J. Effect of three conditions of administration on "trait" and "state" measures of anxiety. *Journal of Consulting and Clinical Psychology*, 1970, **34**, 355–359.

Born, D. G., Gledhill, S. M., & Davis, M. L. Examination performance in lecture-discussion and

personalized instruction courses. *Journal of Applied Behavior Analysis*, 1972, **5**, 45–52.

Husek, T. R., & Alexander, S. The effectiveness of the Anxiety Differential in examination stress situations. *Educationl and Psychological Measurement*, 1963, **23**, 309–318.

Johnston, J. M., & Pennypacker, H. S. A behavioral approach to college teaching. *American Psychologist*, 1971, **26**, 219–244.

Keller, F. S. A personal course in psychology. In R. Ulrich, T. Stachnik, & J. Mabry (Eds.), *The control of human behavior*. Glenview, Ill.: Scott, Foresman, 1966.

McMichael, J. S., & Corey, J. R. Contingency management in an introductory course produces better learning. *Journal of Applied Behavior Analysis*, 1969, **2**, 79–83.

Phares, E. J. Differential utilization of information as a function of internal–external control. *Journal of Personality*, 1968, **36**, 649–662.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, **80**(1, Whole No. 609).

Sarason, I. G. Effect of anxiety and two kinds of motivating instructions on verbal learning. *Journal of Abnormal and Social Psychology*, 1957, **54**, 166–171.

Sheppard, W. C., & MacDermot, H. G. Design and evaluation of a programmed course in introductory psychology. *Journal of Applied Behavior Analysis*, 1970, **3**, 5–11.

Spielberger, C. D. The effects of anxiety on complex learning and academic achievement. In C. D. Spielberger (Ed.), *Anxiety and behavior*. New York: Academic Press, 1966.

Wolfe, R. N. Perceived locus of control and prediction of own academic performance. *Journal of Consulting and Clinical Psychology*, 1972, **38**, 80–83.

(Received April 4, 1974)

(*Continued from p. 910*)

Familiarization of Content and Different Response Modes in Programmed Instruction. Theodore Abramson (Department of Education, Queens College, City University of New York, Flushing, New York 11367) and Edward Kagen.

Concrete and Abstract Thinkers at Three Grade Levels and Their Performance with Complex Concepts. Dean L. Meinke (Department of Educational Psychology, The University of Toledo, 2801 W. Bancroft Street, Toledo, Ohio 43606), Carolyn S. George, and Judith M. Wilkinson.

An Experimental Test of Five Process Models for Subtraction. Lauren Resnick (Learning and D. Center, University of Pittsburgh, Pittsburgh, Pennsylvania 15260), Shirley S. Woods, and Guy S. Groen.

Imagery Rating of 1,000 Frequently Used Words. Barbara W. van der Veur (340 McCracken Hall, Ohio University, Athens, Ohio 45701).

Education and Employment Aspirations of Students: A Probabilistic Approach. Barbara Karmel (School of Business and Technology, Oregon State University, Corvallis, Oregon 97331).

Effects of Word Frequency and Contextual Richness on Children's Word Identification Abilities. P. David Pearson (138 Burton Hall, University of Minnesota, Minneapolis, Minnesota 55455) and Alice Studt.

Establishing Appropriate Purpose for Reading and Its Effect on Flexibility on Reading Rate. S. Jay Samuels (Center for Research in Human Learning, University of Minnesota, Minneapolis, Minnesota 55455) and Patricia Dahl.

Reinforcement and Binet Performance in Disadvantaged Children. Lorene C. Quay (Temple University, College of Education, Philadelphia, Pennsylvania).

Grade Level and Task Difficulty in Learning by The Discovery and Verbal Reception Methods. Evelyn W. Francis (Department of Psychology, Box 3452, New Mexico State University, Las Cruces, New Mexico 88003).

Journal of Educational Psychology
1974, Vol. 66, No. 6, 974-980

EFFECTS OF MODE OF MODELING, MODEL AGE, AND ETHNICITY ON RULE-GOVERNED LANGUAGE BEHAVIORS[1]

JAMES I. GRIESHOP AND MARY B. HARRIS[2]
University of New Mexico

In order to assess the effects of modeling on the performance of syntactic and semantic language behaviors, 208 male and female Anglo and Chicano sixth-grade students were randomly exposed to live, audiotaped, or written models, who were characterized as Anglo or Chicano adults or peers, or to a no-model control group. In the absence of direct or vicarious reinforcement and of instructions to imitate, significant modeling effects for almost all measures in both imitation and generalization phases were found. Within modeling conditions, there were significant tendencies for live and audiotaped models to produce greater imitation of the semantic categories than written models and for adult models to produce greater imitation of the relative clause measure than peer models.

Although the phenomena of modeling and imitative behavior have long been of interest to researchers, the systematic investigation of the effects of social learning variables on rule-governed linguistic and cognitive variables is a much more recent approach (Zimmerman & Rosenthal, 1974). Social learning research in areas such as aggression, altruism, and therapy for phobias has demonstrated that filmed, videotaped, and written models could be effective in altering behaviors, but nevertheless almost all the research on rule-governed cognitive behaviors has beeen done using a live model. Although videotaped (Zimmerman & Dialessi, 1973) and written (Harris & Evans, 1973, 1974) models have been used to influence creative behaviors, there have been no systematic attempts to assess the effects of mode of modeling on any linguistic behaviors. The present study used live, tape-recorded, or written models to see whether their effects would be equivalent or whether the impact of a live model would increase the amount of attention and thus learning and imitation demonstrated by the observer (Bandura, 1971).

A second variable investigated in the present study was the age of the model. Results of research on the influence of peer versus adult models have been contradictory (e.g., Bandura & Kupers, 1964), with some studies finding a greater influence of adult models (Nicolas, McCarter, & Heckel, 1971), some finding peer models more influential (Bandura, Grusec, & Menlove, 1967), and some finding no differences between the two (Malcolm, 1971). The reason for the inconsistency of findings probably lies in the discrepancy between two factors which are both felt to increase the influence of a model: similarity, which is greater for a peer; and expertise or prestige, which is greater for an adult. The present study used adult and

[1] This research was supported in part by a grant from the Student Research Allocations Committee of the University of New Mexico.

[2] The authors wish to express their appreciation to Anselmo Arellano, Jose Armas, Paul Armijo, Pat Baca, Sam Baca, Bobby Bellor, Bert Davis, Andy Donovon, Pat Donovon, David Hegge, David Jaramillo, Andy Joniak, Beatrix Lopez, Richard Rindone, Ernie Sandoval, Douglas Stevenson, Dale Talbot, and Joyce George for their assistance as models, scorers, artist, and secretary.

Requests for reprints should be sent to Mary B. Harris, Department of Educational Foundations, University of New Mexico, Albuquerque, New Mexico 87131.

peer male models to see which of the two would have the greater influence in this instance.

A final variable manipulated was the ethnicity of the model. The models either were (or were presented as, in the written conditions) Chicano or Anglo and identifiable as such by appearance, name, and accent (by name alone in the written model conditions). Three models of each age and ethnic group served in the experiment, to insure that the age and ethnicity variables were not confounded with any unique characteristics of any particular individual. Although no previous studies have investigated the effects of exposure to a Chicano model, investigations of the effects of observing black models and white models by black subjects and white subjects (Breyer & May, 1970; Liebert, Sobol, & Copeman, 1972; Rosenbaum, 1972; Thelen, 1971; Thelen & Freybear, 1971) have produced completely contradictory findings. Garcia and Zimmerman (1972) did find that praise from a Chicano experimenter (not a model) was more reinforcing than praise from an Anglo experimenter. Based on the assumption that similarity to a model would increase imitation (Bandura, 1971), an interaction of ethnic status of model and ethnic status of subject was expected, such that Chicano subjects would imitate Chicano models and Anglo subjects would imitate Anglo models more than vice versa.

Both Chicano and Anglo boys and girls served in the study; no overall ethnic or sex differences were predicted. The dependent measures used in the present study were two semantic valuational categories, two syntactic structures (relative clauses and prepositional phrases), and overall sentence length.

Method

Subjects, Models, and Experimenter

Subjects were 208 sixth-grade students from four elementary schools in Albuquerque, New Mexico. Eight boys and eight girls, half of whom were Chicanos and half were Anglos, were randomly assigned to each of the 12 experimental conditions and to the control group.

Six boys aged 11–12 and six men, half of whom were Chicanos and half were Anglos, were trained and used as models in the live and taped conditions, with approximately equal numbers of subjects exposed to each of the models. In the written condition, the age and names of four real models appeared on the pages handed out to the subjects along with the model's answers.

The experimenter was a 32-year-old male Anglo graduate student.

Stimulus materials. Two parallel but different 12-item sets of line drawings of persons or animals engaged in familiar activities were selected from a larger number of pictures in a pilot study on the basis of eliciting complete sentences from subjects as well as having approximately equivalent numbers of prepositional phrases and equal sentence length. No subjects in the pilot study used any relative clauses or valuational statements. The first 12 pictures served as stimuli for the model and were used in the imitation phase for subjects in the experimental conditions. The other set of pictures was used to assess generalization. Subjects in the control condition simply saw both sets of pictures without one identified as previously having been shown to a model.

Model's statements. A different statement was associated with each of the 12 stimulus pictures seen by a model. Each statement contained a prepositional phrase (e.g., "on a court," "with each other"), a relative clause (e.g., "who is lying," "which carries the baby"), and a valuational–preference belief category (e.g., "Tom believes the best barber—," "the bear doesn't enjoy—"). The mean sentence length for the 12 modeled sentences was 15.25 words.

Procedure

Subjects were run in small groups of 4 to 7, with individuals rather than groups serving as the unit of random assignment. After being escorted by the subject to the experimental site, they were told, "I am interested in how people make up sentences and today I want to find out how you make up sentences." In the live model conditions, a model was also present, and the experimenter told the subjects that another person was going to compose sentences before they did. The model was then asked his name by the experimenter and asked to compose sentences for each of the pictures which were shown by the experimenter to the model and the subjects. After the 12 sentences were composed, the experimenter thanked the model and said to the subjects, "Now that you listened to (name of model) I want you to make up sentences about the same pictures." He then asked each subject to write his name on a sheet of paper and to write his sentences on the sheet. The experimenter then held up the 12 pictures in sequence, waiting until each subject had finished composing a sentence for each picture before proceeding to the next. The papers were then collected and new ones distributed to the subjects,

who were then shown a new set of 12 pictures and asked to make up a sentence about each of them. After this generalization task, the papers were collected, and the subjects were asked not to discuss their experience, and were thanked and dismissed.

Subjects in the other conditions were treated identically except for the presentation of the model. Those in the audiotaped model conditions were told that they would first listen to a tape of another person making up sentences. The tape began with a dialogue between the experimenter and the model, in which the model was asked to state his name and age and asked to compose sentences about the pictures. As the model composed the sentences on tape, the pictures were shown to the subjects by the experimenter. After the model had composed his sentences and been thanked by the experimenter, the tape recorder was turned off and the subjects were treated as in the live model conditions.

Subjects in the written model conditions were told that they would see some sentences written by another person before composing their own. Each subject was then given a set of the 12 stimulus pictures, with the model's sentences written below each picture in handwriting appropriate for a child or adult. The model's name and age appeared on a cover sheet. Since some subjects had difficulty reading the handwriting, the experimenter held up the stimulus cards and read aloud to subjects the sentences which the model had written. The stimulus pictures were then collected, and subjects were asked to compose sentences exactly as in the other conditions.

The control group subjects were treated identically to the experimental subjects with the exception that no presentation or mention of a model was made. They were merely told that the experimenter was interested in how they made up sentences, shown the same pictures as the experimental subjects, and asked to write their sentences.

Scoring. All scoring was done with the identity of the subject, his school, and his experimental condition unknown to the scorers. A random 30% of the papers were independently scored by a second judge, and Pearson product-moment correlations were calculated for each measure in each phase. They ranged from $r = +.92$ for the valuation category–imitation phase to $r = +1.00$ for length in both phases.

There were five dependent measures scored for each of two phases, making a total of 10 measures. The *valuation* category was considered present if a value, preference, or belief attributable to the pictured person or animal was stated in the sentence's predicate, for example, "The boy likes to eat spaghetti." The *other value* category was considered present if any word or phrase expressing a value, preference, or belief was present, for example, "good," "favorite." *Relative clauses* were defined as constructions headed by a relative pronoun (who, that, which) and following a noun phrase, for example, "the man who is brushing his teeth—." *Prepositional phrases* were defined as a phrase consisting of a preposition, an article (with rare exceptions) and a noun, for example, "on a court." *Length* was determined by counting the number of words per response, with contractions scored as single words.

TABLE 1

F Values for A Priori Comparisons of Modeling Versus Control Groups

Phase	Dependent measure				
	Valuational	Other value	Relative clause	Prepositional phrase	Length
Imitation	48.23***	5.62*	6.64**	—	23.82***
Generalization	—	3.12	1.63	13.32***	18.76***

Note.—Not computed, as one-way analysis of variance was not significant.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Results

Control Groups

The scores for the valuational and other value categories, because of the high incidence of zero scores, were analyzed by chi-square analyses. No significant differences were found on either of these measures in either the imitation or generalization phases (all χ^2 values < 2, all $df = 1$, all $ps < .05$). No relative clauses were written by any subject in either phase, making statistical analyses unnecessary. A 2 × 2 (Ethnicity × Sex) analysis of variance was performed on the prepositional phrase and length data. No effect for prepositional phrases was found, but in the generalization phase females wrote significantly more words ($\bar{X} = 88.125$) than males ($\bar{X} = 73.000$, $F = 5.01$, $df = 1/12, p < .05$).

Experimental versus Control Groups

One-way 13-group analyses of variance, collapsing across subject sex and ethnicity revealed significant effects for the valuational category (imitation phase), other value category (both phases), relative clause measure (both phases), prepositional phrase measure (generalization phase), and

TABLE 2

MEAN SCORES FOR 12 SENTENCES FOR MODELING AND CONTROL GROUPS

Phase	Dependent measure				
	Valuational	Other value	Relative clause	Prepositional phrase	Length
Imitation					
All controls	.375	.313	.000	4.000	82.688
All modeling subjects	6.026*	1.167*	1.259*	5.823	115.729*
Live model subjects	7.000[a]	1.016	1.450	5.562	115.344
Taped model subjects	6.719	1.109	1.250	5.594	118.328
Written model subjects	4.359	1.375	1.078	6.312	113.516
Generalization					
All controls	.250	.125	.000	3.438	80.563
All modeling subjects	1.729	.542	.385	6.130*	109.604*
Live model subjects	2.344[a]	.656	.375	5.656	105.500
Taped model subjects	1.797	.359	.203	6.125	111.891
Written model subjects	1.047	.609	.578	6.609	111.422

[a] Live model and taped model subjects higher than written model subjects but not different from each other ($p < .05$).

* Combined modeling subjects higher than combined controls ($p < .05$).

length (both phases) at the $p < .05$ level or beyond (smallest $F = 1.80$, df 12/195). The results of Scheffé a priori comparisons of the 12 modeling groups versus the control group for the measures with significant between-group differences are presented in Table 1. Scores for subjects in the modeling groups were significantly higher than for control group subjects on the valuational (imitation phase), other value (imitation phase), relative clause (imitation phase), prepositional phrase (generalization phase), and length (both phases) measures, indicating a strong modeling effect.

Modeling Groups

Mean scores on the various measures for subjects exposed to different modes of modeling are shown in Table 2. The results of analyses of variance for the main effects of mode of modeling, age of the model, ethnicity of the model, ethnicity of the subject, and sex of subject on all measures for both phases of the experiment are shown in Table 3.[3]

Valuational category. The significant mode of modeling effect in the imitation phase was found by Scheffé a priori comparisons to be due to the fact that subjects exposed to both live ($F = 21.48$, $df = 1/144$, $p < .001$) and taped ($F = 17.15$, $df = 1/144$, $p < .001$) models imitated significantly more than those exposed to a written model, while not differing from each other. In the generalization phase as well, those in the live ($F = 6.37$, $df = 1/144$, $p < .05$) and taped ($F = 3.85$, $df = 1/144$, $p < .05$) model conditions had higher scores than those seeing a written model, while not significantly differing from each other. Significant interactions between model age and subject sex in both the imitation ($F = 6.86$, $df = 1/144$, $p < .05$) and generalization ($F = 4.11$, $df = 1/144$, $p < .05$) phases indicated that male subjects tended to imitate the adult models more than peer models, while female subjects imitated peer models more than adult models.

Other value category. The only meaningful finding revealed by the analyses of variance was a significant effect of subject ethnicity in the generalization phase ($F = 5.65$, $df = 1/144$, $p < .05$), indicating that Anglo subjects gave more other value items ($\bar{X} = .698$) than did Chicano subjects ($\bar{X} = .385$).

Relative clause category. In both the imitation ($F = 4.53$, $df = 1/144$, $p < .05$) and

[3] Several higher order interactions were statistically significant but are not discussed further due to the difficulty of interpreting them meaningfully.

TABLE 3

F VALUES FOR MAIN EFFECTS OF THE ANALYSES OF VARIANCE

Phase	Dependent measure				
	Valuational	Other value	Relative clause	Prepositional phrase	Length
Imitation					
Mode of modeling	12.96***	<1	<1	1.17	<1
Age of model	<1	<1	4.53*	<1	<1
Ethnicity of model	<1	<1	<1	<1	1.06
Ethnicity of subject	<1	<1	<1	2.89	3.61
Sex of subject	<1	<1	<1	<1	<1
Generalization					
Mode of modeling	3.51*	1.97	1.63	1.56	<1
Age of model	<1	2.51	4.36*	<1	<1
Ethnicity of model	<1	<1	1.51	<1	<1
Ethnicity of subject	<1	5.65*	1.82	1.57	1.48
Sex of subject	<1	<1	2.55	<1	3.40

Note. *df* for mode of modeling = 2/144. *df* for all other effects = 1/144.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

generalization ($F = 4.36$, $df = 1/144$, $p < .05$) phases, subjects exposed to an adult model ($\bar{X}$s = 1.615 and .563 for the two phases) wrote more relative clauses than those exposed to a peer ($\bar{X}$s = .979 and .218). In the imitation phase, a significant interaction between mode of modeling and model ethnicity ($F = 3.81$, $df = 2/144$, $p < .05$) indicated that subjects exposed to live ($\bar{X} = 1.813$) and taped Anglo models ($\bar{X} = 1.688$) had higher scores than those subjects exposed to live ($\bar{X} = 1.313$) and taped Chicano models ($\bar{X} = .813$), but that subjects exposed to written Chicano models ($\bar{X} = 1.594$) had higher scores than those who observed written Anglo models ($\bar{X} = 0.563$). In the generalization phase, a significant interaction effect of model age and model ethnicity ($F = 3.86$, $df = 1/144$, $p < .05$) revealed that subjects exposed to Anglo adult models had lower scores ($\bar{X} = .291$) than those subjects exposed to Chicano adult models ($\bar{X} = .833$), but subjects exposed to Anglo peer models had higher scores ($\bar{X} = .271$) than those subjects who observed Chicano peer models ($\bar{X} = .146$).

Prepositional phrase category. The only significant effect was an interaction between model age and model ethnicity in the generalization phase ($F = 5.70$, $df = 1/144$, $p < .05$), which showed that subjects observing Anglo adult models ($\bar{X} = 5.729$) had lower scores than those subjects who observed Chicano adult models ($\bar{X} = 6.813$), while subjects exposed to Anglo peer models ($\bar{X} = 6.333$) had higher scores than those subjects exposed to Chicano peer models ($\bar{X} = 5.479$).

Length category. Although no significant effects were found in the imitation phase, an interaction between mode of modeling and model age ($F = 4.569$, $df = 2/144$, $p < .05$) in the generalization phase revealed that subjects exposed to live ($\bar{X} = 110.625$) and written adult models ($\bar{X} = 115.750$) wrote more words than those subjects exposed to live ($\bar{X} = 100.375$) and written peer models ($\bar{X} = 107.094$), while subjects hearing taped adult models wrote fewer words ($\bar{X} = 103.125$) than those subjects who heard the tape peer models ($\bar{X} = 118.250$).

DISCUSSION

The results of this study indicated that sixth-grade students, with neither explicit or implicit instructions to imitate nor reinforcement, were able to abstract rules governing the use of modeled sentences and subsequently to use those rules to generate new

sentences in response to novel stimuli. In particular, the students exposed to a model abstracted the rules governing the valuational categories and relative clauses and then used those rules in both imitative and novel tasks to write sentences which expressed the modeled values and syntactic structures, whereas the no-model control students did not write sentences with the same number of semantic or syntactic structures. The modeling effect, since it occurred without instructions or reinforcement, provides further evidence that imitation of models can occur without them (Bandura 1971).

The results, however, only partially clarified the role and influence of specified attentional variables within the social learning theory paradigm. The mode of modeling and the age of the model had some effects upon the abstraction and imitation of the modeled language categories, both directly and in interaction with other factors, particularly the sex of the observer. The ethnicity of the model, like that of the observer, had little effect upon the modeling phenomenon. The latter observation tends to corroborate the conclusion of Harris and Hassemer (1972) that the language (Spanish or English) used by the model had no effect. The sex of the observer alone had no apparent effect upon the outcomes, although it did seem to interact significantly with age of model.

Much like the results of the Rosenthal, Zimmerman, and Durning (1970) study, no mimicry or precise imitation of the models and the modeled sentences were found. This lack of mimicry further supports the social learning theory belief that modeling procedures can be used to generate new and creative responses which conform to the modeled rule-governed behaviors. Another interesting feature was the use of the other value category by subjects exposed to a model, although only three of the model's sentences embodied this category. This fact suggests that the students in the present study, similar to those subjects used by Odum, Liebert, and Hill (1968) and Liebert, Odum, Hill, and Huff (1969), seemed to employ problem-solving strategies to analyze the modeled sentences, and generalized their behavior beyond the specific categories modeled.

Methodological Implications

Several methodological innovations incorporated in this study have special significance for social learning theory and for the practical application of modeling procedures.

The fact that 12 different individuals served as the models for this study negates the possibility that modeling effects may be due to some idiosyncracy of the model used. This fact is particularly meaningful since the majority of modeling studies have used only a single model, and it implies that in an applied setting any of a number of potential models can be used with a degree of effectiveness.

A second innovation was the use of peers as models. The finding that the scores of subjects (with the exception of the scores for the relative clause measure) exposed to peer models were nearly equal to those of subjects observing adult models suggests that peers may be effectively used as models displaying rule-governed behaviors for sixth-grade children, although it is possible that with complex grammatical constructions, such as relative clauses, adult models may be more appropriate.

The small groupings of students used in this study indicate the modeling effects for rule-governed behaviors occur in group situations similar to classroom settings. This finding replicates that made by Rosenthal and Carroll (1972), and has obvious practical significance.

The fact that the subjects were required to compose their responses in written form without any word lists as cues also distinguished this study from other modeling studies. The results clearly demonstrated that the specified rules and behaviors could be orally modeled, transmitted and abstracted, and finally converted by the subjects into written form. The procedural innovation and consequent finding imply that the procedure could well be incorporated as a classroom technique for increasing the performance of rule-governed

language behaviors. Moreover, the results imply that for some categories of behavior, taped and even written examples can be used to supplement the teacher as a model.

REFERENCES

Bandura, A., *Social learning theory*. New York: General Learning Press, 1971.

Bandura, A., Grusec, J. E., & Menlove, F. L. Some social determinants of self-monitoring reinforcement systems. *Journal of Personality and Social Psychology*, 1967, **5**, 449–455.

Bandura, A., & Kupers, C. J. Transmission of patterns of self-reinforcement through modeling. *Journal of Abnormal and Social Psychology*, 1964, **69**, 1–9.

Breyer, N. L., & May, J. G. Effect of sex and race of the observer and model on imitation learning. *Psychological Reports*, 1970, **27**, 639–646.

Garcia, A. B., & Zimmerman, B. J. The effect of examiner ethnicity and language on the performance of bilingual Mexican-American first graders. *Journal of Social Psychology*, 1972, **87**, 3–11.

Harris, M. B., & Evans, R. C., Models and creativity. *Psychological Reports*, 1973, **33**, 763–769.

Harris, M. B., & Evans, R. C. The effects of models and instructions on creative responses. *Journal of Psychology*, 1974, **86**, 123–130.

Harris, M. B., & Hassemer, W. G. Some factors affecting the complexity of children's sentences: The effects of modeling, age, sex and bilingualism. *Journal of Experimental Child Psychology*, 1972, **13**, 447–455.

Liebert, R. M., Odom, R. D., Hill, I. H., & Huff, R. L. Effects of age and rule familiarity on the production of modeled language constructions. *Developmental Psychology*, 1969, **1**, 108–112.

Liebert, R. M., Sobol, M. P., & Copeman, C. D. Effects of vicarious consequences and race of model upon imitative performance by black children. *Developmental Psychology*, 1972, **6**, 453–456.

Malcolm, D. B. The relative efficacy of adult and peer models on children's moral judgments. *Dissertation Abstracts International*, 1971, **32**, 246A.

Nicholas, K. B., McCarter, R. E., & Heckel, R. V. Imitation of adult and peer television models by white and Negro children. *Journal of Social Psychology*, 1971, **85**, 317–318.

Odom, R. D., Liebert, R. M., & Hill, J. H. The effects of modeling cues, reward and attentional set on the production of grammatical and ungrammatical syntactic constructions. *Journal of Experimental Child Psychology*, 1968, **6**, 131–140.

Rosenbaum, E. Effects of race of observer, examiner, and model on imitation of a school-like task. *Dissertation Abstracts International*, 1972, **32**, 4428A.

Rosenthal, T. L., & Carroll, W. R. Factors in vicarious modification of complex grammatical parameters. *Journal of Educational Psychology*, 1972, **63**, 174–178.

Rosenthal, T. L., Zimmerman, B. J., & Durning, K. Observationally induced changes in children's interrogative classes. *Journal of Personality and Social Psychology*, 1970, **16**, 681–688.

Thelen, M. H. The effects of subject race, model race, and vicarious praise on vicarious learning. *Child Development*, 1971, **42**, 972–977.

Thelen, M. H., & Frybear, J. L. Effect of observer and model race on the imitation of standards of self-reward. *Developmental Psychology*, 1971, **5**, 133–135.

Zimmerman, B. J., & Dialessi, F. Modeling influences on children's creative behavior. *Journal of Educational Psychology*, 1973, **65**, 127–134.

Zimmerman, B. J., & Rosenthal, T. L. Observational learning of rule-governed behavior by children. *Psychological Bulletin*, 1974, **81**, 29–42.

(Received April 15, 1974)

Journal of Educational Psychology

JOANNA WILLIAMS, EDITOR
Teachers College, Columbia University and University of Pennsylvania

SAMUEL BALL, ASSOCIATE EDITOR
Educational Testing Service

ADVISORY EDITORS

PAUL R. AMMON, *University of California, Berkeley*
JERE BROPHY, *University of Texas at Austin*
C. VICTOR BUNDERSON, *Brigham Young University*
LEONARD S. CAHEN, *Educational Testing Service and Far West Laboratory for Educational Research and Development*
LOIS-ELLIN DATTA, *National Institute of Education*
FREDERICK B. DAVIS, *University of Pennsylvania*
FRANCIS J. DIVESTA, *Pennsylvania State University*
ROBERT EBEL, *Michigan State University*
DAVID ELKIND, *University of Rochester*
LAWRENCE T. FRASE, *Bell Laboratories*
N. L. GAGE, *Stanford University*
ROBERT GAGNÉ, *Florida State University*
ROBERT GLASER, *University of Pittsburgh*
MIRIAM L. GOLDBERG, *Teachers College, Columbia University*
EDMUND W. GORDON, *Teachers College, Columbia University*
VERNON C. HALL, *Syracuse University*
ARTHUR R. JENSEN, *University of California, Berkeley*
PHYLLIS A. KATZ, *City University of New York, Graduate School and University Center*
HERBERT J. KLAUSMEIER, *University of Wisconsin—Madison*
ROBERT E. LANA, *Temple University*
GERALD S. LESSER, *Harvard University*
JOEL R. LEVIN, *University of Wisconsin—Madison*
FREDERICK J. MCDONALD, *Educational Testing Service*
WILBERT J. MCKEACHIE, *University of Michigan*
FRANK B. MURRAY, *University of Delaware*
LAUREN B. RESNICK, *University of Pittsburgh*
WILLIAM D. ROHWER, JR., *University of California, Berkeley*
ROBERT ROSENTHAL, *Harvard University*
ERNST Z. ROTHKOPF, *Bell Laboratories*
S. JAY SAMUELS, *University of Minnesota*
SEYMOUR B. SARASON, *Yale University*
MELVYN I. SEMMEL, *Indiana University*
IRVING SIGEL, *Educational Testing Service*
HARRY SILBERMAN, *University of California, Los Angeles*
RICHARD E. SNOW, *Stanford University*
GEORGE G. THOMPSON, *Ohio State University*

VOLUME 66, 1974

HAROLD P. VAN COTT
Managing Editor

ANITA DEVIVO
Executive Editor and General Manager

LINDA B. BARODY
Supervising Technical Editor

SHERRI RICARDO
Technical Editor

LYNN D. SPIER
Acting Manager, Publications Business Service

ANNE REDMAN
Subscription Manager

Published bimonthly by:

The American Psychological Association, Inc.
1200 Seventeenth Street, N.W.
Washington, D. C. 20036

Copyright © 1974 by the American Psychological Association, Inc.

Announcement

In addition to the members of the Advisory Board, the *Journal of Educational Psychology* asks other psychologists to aid in the review of manuscripts. In this way we have been able to achieve a broader review base. The Editors wish to thank the following people for their generous assistance to the Journal.

Marjorie Arnold, *Rutgers University, New Brunswick*
Leonard Blackman, *Columbia University, Teachers College*
Marilynn Brewer, *Loyola University*
Betty Bryant, *Educational Testing Service*
Richard Christie, *Columbia Universi*
William Coffman, *University of Iowa*
Frank Costin, *University of Illinois*
Richard Cox, *University of Pittsburgh*
Carol Dwyer, *Educational Testing Service*
Ruth Ekstrom, *Educational Testing Service*
Edgar Epps, *University of Chicago*
Stanford Ericksen, *University of Michigan*
Gerard C. Fanelli, *Newark State College*
John Feldhusen, *Purdue University*
John Follman, *University of South Florida*
Garlie Forehand, *Carnegie Institute of Technology*
Norman Fredericksen, *Educational Testing Service*
John Fremer, *Educational Testing Service*
Anthony Grasha, *University of Cincinnati*
Richard Hansen, *City College of City University of New York*
Richard Jaeger, *University of South Florida*
Jerome Kagan, *Harvard University*
Martin Kaufman, *U. S. Office of Education*
Jacob Kounin, *Wayne State University*
Gerald Kulm, *Purdue University*

CONTENTS OF VOLUME 66

Abelson, Willa D., Zigler, Edward, and DeBlasi, Cheryl L. Effects of a four-year follow through program on economically disadvantaged children 756
Adelman, Howard. *See* Feshbach, Seymour.
Airasian, Peter W. *See* Bart, William M.
Allen, George J., Giat, Laura, and Cherney, Roger J. Locust of control, test anxiety, and student performance in a personalized instruction course 968
Alpert, Judith Landon. Teacher behavior across ability groups: A consideration of the mediation of Pygmalion effects 348
Alston, Herbert L. *See* Doughtie, Eugene B.
Anderson, O. Roger. *See* Browne, Robert J.
Anderson, Richard C. Concretization and sentence learning 179
Archwamety, Teara. *See* Samuels, S. Jay.
Armel, Sandra. *See* Taffel, Suzanne Johnson.
Asher, Steven R., and Markell, Richard A. Sex differences in comprehension of high- and low-interest reading material 680
Austin, Gilbert R., and Postlethwaite, T. Neville. Cognitive results based on different ages of entry to school: A comparative study 857
Babad, Elisha Y., and Budoff, Milton. Sensitivity and validity of learning-potential measurement in three levels of ability 439
Baker, Phoebe K. *See* Davidson, Robert E.
Bart, William M., and Airasian, Peter W. Determination of the ordering among seven Piagetian tasks by an ordering-theoretic method 277
Baumeister, Alfred A. Serial memory span thresholds of normal and mentally retarded children 889
Berliner, David C. *See* Shavelson, Richard J.
Bernbaum, Marcia, Goodnow, Jacqueline, and Lehman, Elyse. Relationships among perceptual–motor tasks: Tracing and copying 731
Bersoff, Donald N. *See* Byalick, Robert.
Billington, Marjorie J. *See* Rothkopf, Ernst Z.
Boker, John R. Immediate and delayed retention effects of interspersing questions in written instructional passages 96
Bower, Gordon H. Selective facilitation and interference in retention of prose 1
Block, James H., and Tierney, Michael L. An exploration of two correction procedures used in mastery learning approaches to instruction 962
Brandt, Larry J., and Hayden, Mary Ellen. Male and female teacher attitudes as a function of students' ascribed motivation and performance levels 309
Brassell, William R. *See* Rollins, Howard A.
Bridgeman, Brent. Effects of test score feedback on immediately subsequent test performance 62
Brophy, Jere. *See* Willis, Sherry.
Brophy, Jere E. *See* Good, Thomas L.
Brophy, Jere E. *See* Veldman, Donald J.
Broverman, Donald M. *See* Mackenberg, Edmund J.
Brown, Bobby R. *See* Duchastel, Philippe C.
Browne, Robert J., and Anderson, O. Roger. Lesson kinetic structure analysis as related to pupil awareness and achievement 864
Budoff, Milton. *See* Babad, Elisha Y.
Burkett, James R. *See* Siegel, Arthur I.
Byalick, Robert, and Bersoff, Donald N. Reinforcement practices of black and white teachers in integrated classrooms 473
Capps, Carolyn. *See* Spring, Carl.
Carrier, Neil A., Howard, George S., and Miller, William G. Course evaluation: When? . 609
Chalmers, Douglas K., and Rosenbaum, Milton E. Learning by observing versus learning by doing 216
Cherney, Roger J. *See* Allen, George J.
Cobb, Joseph A. *See* Hops, Hyman.
Coke, Esther U. The effects of readability on oral and silent reading rates 406
Connelly, Francis. *See* Nord, Walter R.

Craighead, W. Edwards. *See* Mercatoris, Michael.
Crano, William D. *See* Schmidt, Frank L.
Crouse, James H. Acquisition of college course material under conditions of repeated testing 367
Crouse, James H. *See* Price, Jay R.
Crowl, Thomas K., and MacGinitie, Walter H. The influence of students' speech characteristics on teachers' evaluations of oral answers 304
Cunningham, Donald J. *See* Garrett, Candace S.
Dahl, Patricia, *See* Samuels, S. Jay.
Daignault, George. *See* Nord, Walter R.
Davidson, Robert E., Perry, Sherry A. B., and Baker, Phoebe K. Unfamiliar stimulus terms in children's paired-associate learning 580
Dawis, René V. *See* Whitely, Susan E.
DeBlasi, Cheryl L. *See* Abelson, Willa D.
Denney, Douglas R. Relationship of three cognitive style dimensions to elementary reading abilities 702
Desberg, Peter. *See* Marsh, George.
Devries, David L., and Edwards, Keith J. Student teams and learning games: Their effects on cross-race and cross-sex interaction 741
Diaz-Guerrero, Rogelio, and Holtzman, Wagne H. Learning by televised "Plaza Sesamo" in Mexico 632
DiVesta, Francis J. *See* Rickards, John P.
Divine-Hawkins, Patricia. *See* Levin, Joel R.
Doughtie, Eugene B., Wakefield, James A., Jr., Sampson, Raymond N., and Alston, Herbert L. A statistical test of the theoretical model for the representational level of the Illinois Test of Psycholinguistic Ability 410
Doyle, Walter, and Redwine, Judith McNally. Effect of intent-action discrepancy and student performance feedback on teacher behavior change 750
Duc Tuan, Nguyen. *See* Rychlak, Joseph F.
DuCette, Joseph. *See* Wolk, Stephen.
Duchastel, Philippe C., and Brown, Bobby R. Incidental and relevant learning with instructional objectives 481
Duell, Orpha K. Effect of type of objective, level of test questions, and the judged importance of tested materials upon posttest performance 225
Dusek, Jerome B. *See* O'Connell, Edward J.
Dwyer, Carol Anne. Influence of children's sex role standards on reading and arithmetic achievement 811
Dyer, James W., and Kulhavy, Raymond W. Sequence effects and reading time in programmed learning 57
Dyer, Jean L., and Miller, Louise B. Notes on Crano, Kenny, and Campbell's "Does intelligence cause achievement?" 49
Edwards, Keith J. *See* Devries, David L.
Ehrenpreis, Walter, and Scandura, Joseph M. The algorithmic approach to curriculum construction: A field test in mathematics 491
Ehri, Linnea C., and Muzio, Irene M. Cognitive style and reasoning about speed 569
Elmore, Patricia B., and LaPointe, Karen A. Effects of teacher sex and student sex on the evaluation of college instructors 386
Elton, Charles F., and Rose, Harriett A. Prediction of productivity from personality test scores 424
Evans, Franklin R. *See* Frederiksen, Norman.
Farwell, Lawrence Kent. *See* Marsh, George.
Feather, N. T. Coeducation, values, and satisfaction with school 9
Felixbrod, Jeffrey J., and O'Leary, K. Daniel. Self-determination of academic standards by children: Toward freedom from external control 845
Fenner, Diane. *See* Winchel, Ronald.
Feshbach, Seymour, and Adelman, Howard. Remediation of learning problems among the disadvantaged 16
Frederiksen, Janet D., and Rohwer, William D., Jr. Elaborative prompt effects in children's paired-associate learning: Design and population comparisons 83
Frederiksen, Norman, and Evans, Franklin R. Effects of models of creative performance on ability to formulate hypotheses 67
Freijo, Tom D. *See* Jaeger, Richard M.
French-Lazovik, Grace. Predictability of students' evaluations of college teachers from component ratings 373
Fleishman, Allen. *See* Humphreys, Lloyd G.

Garrett, Candace S., and Cunningham, Donald J. Effects of vicarious consequences and model and experimenter sex on imitative behavior in first-grade children 940
Giat, Laura. *See* Allen, George J.
Goldberg, Frank. Effects of imagery on learning incidental material in the classroom 233
Good, Thomas L., and Brophy, Jere E. Changing teacher and student behavior: An empirical investigation 390
Goodnow, Jacqueline. *See* Bernbaum, Marcia.
Goulet, L. R., Williams, Kerry G., and Hay, Carl M. Longitudinal changes in intellectual functioning in preschool children: Schooling- and age-related effects 657
Gray, Jerry L. *See* Robinson, Jack E.
Grieshop, James I., and Harris, Mary B. Effects of mode of modeling, model age, and ethnicity on rule-governed language behaviors 974
Gump, Paul V. *See* Kounin, Jacob S.
Guttmann, Joseph. *See* Levin, Joel R.
Hall, Vernon C., and Russell, William J. C. Multitrait-multimethod analysis of conceptual tempo 932
Hansen, Joe B. Effects of feedback, learner control, and cognitive abilities on state anxiety and performance in a computer-assisted instruction task 247
Harris, Chester W., and McArthur, David L. Another view of the relation of environment to mental abilities 457
Harris, Mary B. *See* Grieshop, James I.
Hay, Carl M. *See* Goulet, L. R.
Hayden, Mary Ellen. *See* Brandt, Larry J.
Heads, Tom B. *See* Marholin, David, II.
Herson, Phyllis F. Biasing effects of diagnostic labels and sex of pupil on teachers' views of pupils' mental health 117
Hiller, Jack H. Learning from prose text: Effects of readability level, inserted question difficulty, and individual differences 202
Holtzman, Wagne H. *See* Diaz-Guerrero, Rogelio.
Hops, Hyman, and Cobb, Joseph A. Initial investigations into academic survival-skill training, direct instruction, and first-grade achievement 548
House, Ernest R. *See* Steele, Joe M.
Howard, George S. *See* Carrier, Neil A.
Humphreys, Lloyd G., and Fleishman, Allen. Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables 464
Hyde, Elizabeth M. *See* Jorgenson, Gerald W.
Jaeger, Richard M., and Freijo, Tom D. Some psychometric questions in the evaluation of professors 416
Jensen, Arthur R. Interaction of Level I and Level II abilities with race and socioeconomic status 99
Jensen, Mary, and Rosenfeld, Lawrence B. Influence of mode of presentation, ethnicity, and social class on teachers' evaluations of students 540
Johnson, James. *See* Saltz, Eli.
Johnson, Roger A. Thinking of two economic levels of elementary school children 530
Johnson, Ronald E. Abstractive processes in the remembering of prose 772
Jones, Keith. *See* Perry, Raymond P.
Jorgenson, Gerald W., and Hyde, Elizabeth M. Auditory–visual integration and reading performance in lower-social-class children 718
Kaplan, Robert. Effects of learning prose with part versus whole presentations of instructional objectives 787
Kaplan, Robert, and Rothkopf, Ernst Z. Instructional objectives as directions to learners: Effect of passage length and amount of objective-relevant content 448
Kaplan, Robert, and Simmons, Francine G. Effects of instructional objectives used as orienting stimuli or as summary/review upon prose learning 614
Katz, Phyllis A., and Zalk, Sue Rosenberg. Doll preferences: An index of racial attitudes 663
Kaye, Kenneth. *See* Wainer, Howard.
Kerst, Stephen M. *See* Levin, Joel R.
Klaiber, Edward L. *See* Mackenberg, Edmund J.
Klausmeier, Herbert J. *See* Nelson, Gordon K.
Kogan, Nathan, and Pankove, Ethel. Long-term predictive validity of divergent-thinking tests: Some negative evidence 802
Kohn, Martin, and Rosman, Bernice L. Social-emotional, cognitive, and demographic determinants of poor school achievement: Implications for a strategy of intervention 267

Kounin, Jacob S., and Gump, Paul V. Signal systems of lesson settings and the task-related behavior of preschool children 554
Kulhavy, Raymond W. *See* Dyer, James W.
Kulhavy, Raymond W. *See* Swenson, Ingrid.
LaForge, Rolfe. *See* Yee, Leland Y.
LaPointe, Karen A. *See* Elmore, Patricia B.
Lautman, Martin R. *See* Siegel, Arthur I.
Lehman, Elyse. *See* Bernbaum, Marcia.
Lesgold, Alan M. Variability in children's comprehension of syntactic structures 333
Levin, Joel R. *See* Varley, William H.
Levin, Joel R., Divine-Hawkins, Patricia, Kerst, Stephen M., and Guttmann, Joseph. Individual differences in learning from pictures and words: Development and application of an instrument 296
Linke, Russell D. Influence of cultural background on hierarchical learning 911
Loeding, David. *See* Shavelson, Richard J.
Loftus, Elizabeth F., and Loftus, Geoffrey R. Changes in memory structure and retrieval over the course of instruction 315
Loftus, Geoffrey R. *See* Loftus, Elizabeth F.
MacGinitie, Walter H. *See* Crowl, Thomas K.
Mackenberg, Edmund J., Broverman, Donald M., Vogel, William, and Klaiber, Edward L. Morning-to-afternoon changes in cognitive performances and in the electroencephalogram 238
MacMillan, Donald L., and Wright, Deborah L. Outerdirectedness in children of three ages as a function of experimentally induced success and failure 919
Marholin, David, II, McInnis, Elizabeth T., and Heads, Tom, B. Effect of two free-time reinforcement procedures on academic performance in a class of behavior problem children 872
Marjoribanks, Kevin. Another view of the relation of environment to mental abilities: A reply 460
Markell, Richard A. *See* Asher, Steven R.
Marsh, George, Desberg, Peter, and Farwell, Lawrence Kent. Stimulus and response variables in children's learning of grapheme–phoneme correspondences 112
Martin, Roy. *See* Meyers, Joel.
Martuza, Victor R. *See* Price, Jay R.
Marwit, Samuel J., and Neumann, Gail. Black and white children's comprehension of standard and nonstandard English passages 329
Mattsson, Kenneth D. Personality traits associated with effective teaching in rural and urban secondary schools 123
Mayer, Richard E. Acquisition processes and resilience under varying testing conditions for structually different problem-solving procedures 644
McArthur, David L. *See* Harris, Chester W.
McCandless, Boyd R. *See* Pusser, H. Ellison.
McCandless, Boyd R. *See* Rollins, Howard A.
McInnis, Elizabeth T. *See* Marholin, David, II.
Mercatoris, Michael, and Craighead, W. Edward. Effects of nonparticipant observation on teacher and pupil classroom behavior 512
Merrill, Paul F. Effects of the availability of objectives and/or rules on the learning process 534
Meyers, Joel, and Martin, Roy. Relationships of state and trait anxiety to concept-learning performance 33
Miller, Louise B. *See* Dyer, Jean L.
Miller, William G. *See* Carrier, Neil A.
Mischel, Harriet N. Sex bias in the evaluation of professional achievements 157
Muzio, Irene M. *See* Ehri, Linnea C.
Nelson, Gordon K., and Klausmeier, Herbert J. Classificatory behaviors of low-socio-economic-status children 432
Nelson, Rosemery O., and Wein, Kenneth S. Training letter discrimination by presentation of high-confusion versus low-confusion alternatives 926
Neumann, Gail. *See* Marwit, Samuel J.
Newman, John E. Sex differences in the organizational assimilation of beginning graduate students in psychology 129
Niemi, R. Ronald. *See* Perry, Raymond P.
Nord, Walter R., Connelly, Francis, and Daignault, George. Locus of control and aptitude test scores as predictors of academic achievement 956

O'Connell, Edward J., Dusek, Jerome B., and Wheeler, Richard J. A follow-up study of teacher expectancy effects 325
O'Leary, K. Daniel. *See* Felixbrod, Jeffrey J.
O'Leary, K. Daniel. *See* Taffel, Suzanne Johnson.
Pambookian, Hagop S. Initial level of student evaluation of instruction as a source of influence on instructor change after feedback 52
Pankove, Ethel. *See* Kogan, Nathan.
Peeck, J. Retention of pictorial and verbal content of a text with illustrations 880
Perry, Raymond P., Niemi, R. Ronald, and Jones, Keith. Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance 851
Perry, Sherry A. B. *See* Davidson, Robert E.
Pomerantz, Michael. *See* Schultz, Charles B.
Postlethwaite, T. Neville. *See* Austin, Gilbert R.
Price, Jay R., Martuza, Victor R., and Crouse, James H. Construct validity of test items measuring acquisition of information from line graphs 152
Pusser, H. Ellison, and McCandless, Boyd R. Socialization dimensions among inner-city five-year-olds and later school success: A follow-up 285
Ravitch, Michael M. *See* Shavelson, Richard J.
Redwine, McNally Judith. *See* Doyle, Walter.
Resnick, Lauren B. *See* Rosenthal, Daniel J. A.
Rhetts, John E. Task, learner, and treatment variables in instructional design 339
Richards, James M., Jr. "Environments" of British Commonwealth universities 572
Rickards, John P., and DiVesta, Francis J. Type and frequency of questions in processing textual material 354
Robinson, Jack E., and Gray, Jerry L. Cognitive style as a variable in school learning ... 793
Rohwer, William D., Jr. *See* Frederiksen, Janet D.
Rollins, Howard A., McCandless, Boyd R., Thompson, Marion, and Brassell, William R. Project Success Environment: An extended application of contingency management in inner-city schools 167
Rose, Harriett, A. *See* Elton, Charles F.
Rosenbaum, Milton E. *See* Chalmers, Douglas K.
Rosenfeld, Lawrence B. *See* Jensen, Mary.
Rosenthal, Daniel J. A., and Resnick, Lauren B. Children's solution processes in arithmetic word problems 817
Rosman, Bernice L. *See* Kohn, Martin.
Rothkopf, Ernst Z. *See* Kaplan, Robert.
Rothkopf, Ernst Z., and Billington, Marjorie J. Indirect review and priming through questions 669
Rowe, Edward J. *See* Skanes, Graham R.
Rubin, Donald B. Estimating causal effects of treatments in randomized and nonrandomized studies 688
Russell, William J. C. *See* Hall, Vernon C.
Rychlak, Joseph F., Duc Tuan, Nguyen, and Schneider, William E. Formal discipline revisited: Affective assessment and nonspecific transfer 139
Salomon, Gavriel. Internalization of filmic schematic operations in interaction with learners' aptitudes 499
Saltz, Eli, and Johnson, James. Training for thematic-fantasy play in culturally disadvantaged children: Preliminary results 623
Sampson, Raymond N. *See* Doughtie, Eugene B.
Samuels, S. Jay, and Turnure, James E. Attention and reading achievement in first-grade boys and girls 29
Samuels, S. Jay, Dahl, Patricia, and Archwamety, Teara. Effect of hypothesis/test training on reading skill 835
Scandura, Joseph M. *See* Ehrenpreis, Walter.
Schmidt, Frank L., and Crano, William D. A test of the theory of fluid and crystallized intelligence in middle- and low-socioeconomic-status children: A cross-lagged panel analysis 255
Schneider, William E. *See* Rychlak, Joseph F.
Schultz, Charles B. Information seeking following the confirmation or contradiction of beliefs 903
Schultz, Charles B., and Pomerantz, Michael. Assessment of motive to succeed and probability of success 599
Severson, Roger A. *See* Sewell, Travor E.
Severson, Roger A. *See* Varley, William H.

Sewell, Trevor E., and Severson, Roger A. Learning ability and intelligence as cognitive predictors of achievement in first-grade black children 948
Shannon, Elizabeth. *See* Skanes, Graham R.
Shavelson, Richard J., Berliner, David C., Ravitch, Michael M., and Loeding, David. Effects of position and type of question on learning from prose material: Interaction of treatments with individual differences 40
Shaver, Phillip. *See* Winchel, Ronald.
Siegel, Arthur I., Lautman, Martin R., and Burkett, James R. Reading grade level adjustment and auditory supplementation as techniques for increasing textual comprehensibility 895
Simmons, Francine, G. *See* Kaplan, Robert.
Skanes, Graham R. *See* Sullivan, Arthur M.
Skanes, Graham R., Sullivan, Arthur M., Rowe, Edward J., and Shannon, Elizabeth. Intelligence and transfer: Aptitude by treatment interactions 563
Spring, Carl, and Capps, Carolyn. Encoding speed, rehearsal, and probed recall of dyslexic boys 780
Steele, Joe M., Walberg, Herbert J., and House, Ernest R. Subject areas and cognitive press 363
Stuart, Richard B. Teaching facts about drugs: Pushing or preventing? 189
Sullivan, Arthur M. *See* Skanes, Graham R.
Sullivan, Arthur M., and Skanes, Graham R. Validity of student evaluation of teaching and the characteristics of successful instructors 584
Swenson, Ingrid, and Kulhavy, Raymond W. Adjunct questions and the comprehension of prose by children 212
Taffel, Suzanne Johnson, O'Leary, K. Daniel, and Armel, Sandra. Reasoning and praise: Their effects on academic behavior 291
Thompson, Marion. *See* Rollins, Howard A.
Tierney, Michael L. *See* Block, James H.
Turnure, James E. *See* Samuels, S. Jay.
Varley, William H., Levin, Joel R., Severson, Roger A., and Wolff, Peter. Training imagery production in young children through motor involvement 262
Veldman, Donald J., and Brophy, Jere E. Measuring teacher effects on pupil achievement 319
Vogel, William. *See* Mackenberg, Edmund J.
Voss, James F. Acquisition and nonspecific transfer effects in prose learning as a function of question form 736
Wainer, Howard, and Kaye, Kenneth. Multidimensional scaling of concept learning in an introductory course 591
Wakefield, James A., Jr. *See* Doughtie, Eugene B.
Walberg, Herbert J. *See* Steele, Joe M.
Walker, Basil S. Effects of inserted questions on retroactive inhibition in meaningful verbal learning 486
Wein, Kenneth S. *See* Nelson, Rosemery O.
Wheeler, Richard J. *See* O'Connell, Edward J.
Whitely, Susan E., and Dawis, René V. Effects of cognitive intervention on latent ability measured from analogy items 710
Williams, Kerry G. *See* Goulet, L. R.
Willis, Sherry, and Brophy, Jere. Origins of teachers' attitudes toward young children 520
Winchel, Ronald, Fenner, Diane, and Shaver, Phillip. Impact of coeducation on "Fear of Success" imagery expressed by male and female high school students 726
Wolk, Stephen, and DuCette, Joseph. Monetary incentive effects upon incidental learning during an instructional task 90
Wolff, Peter. *See* Varley, William H.
Woodson, M. I. Chas. E. Seven aspects of teaching concepts 184
Wright, Deborah L. *See* MacMillan, Donald L.
Yee, Leland Y., and LaForge, Rolfe. Relationship between mental abilities, social class, and exposure to English in Chinese fourth graders 826
Zigler, Edward. *See* Abelson, Willa D.
Zalk, Sue Rosenberg. *See* Katz, Phyllis A.

SMITH DELIVERS!

Smith delivers a solidly applied treatment of the multidimensional aspects of educational psychology. Focusing all concepts on the teacher, the learner, and the teaching environment, the book bridges the gap between learning theories and teaching processes in a way that is applicable to any subject at any level.

Challenging, point for point, the book provides up-to-date coverage of such topics as: instructional technology, theory and science of instruction, humanistic psychology, learning disabilities, language development, group dynamics, evaluation and many, many more areas of current importance and concern. The wide range of relevant topics covered makes this book a most comprehensive treatment of all traditional and modern aspects of educational psychology.

New 1975!
In a lively, clear and straightforward manner, the style of the book involves the reader.
— Involves the reader with many concrete, practical applications, examples and case studies to help smooth the transition from theory to practice.
— Involves the reader with relevant illustrative material including cartoons, charts, and photographs all presented in a modern, readable, two-color format.
— Involves the reader with thought provoking and stimulating chapter questions and exercises and innovative classroom projects and activities.

Educational Psychology and its Classroom Applications

M. Daniel Smith

University of New Hampshire

Smith delivers a complete teaching package which includes:
☐ **Student Study Guide** containing behavioral objectives, pre-tests, and post-tests.
☐ **Instructor's Supplement** featuring a variety of alternative teaching strategies and course formats in which the book can be used. The guide also suggests ways of motivating students and approaches to assessing student understanding.
☐ **A Manual of Test Items** including the behavioral objectives the tests are designed to reinforce.
January 1975, paperbound, est. 600 pp.

Allyn and Bacon, Inc.

College Division, Department 893
470 Atlantic Ave., Boston, MA 02210



Behavior Dynamics is a broad, general survey of the psychological aspects of human behavior that are reflected in teaching, learning, and growth. Advocating a humanistic perspective, Hamachek develops a psychological framework for understanding and interpreting behavior; examines physical, psychological, and intellectual growth processes as related to developmental outcomes and behavioral consequences; looks at the dynamics of "good" teaching in terms of self-understanding and the ability to make instruction relevant and meaningful; studies learning processes and relationships to motivational inputs, self-concept variables, and measurement procedures; and presents an overview of ways of understanding and managing classroom groups.

February 1975, paperbound, est. 672 pp.

Behavior Dynamics in Teaching, Learning and Growth

New 1975!

Hamachek's dynamic writing style is both clear and informal, personalizing the solid informational base of clinical evidence and empirical research with verbal and visual illustrations and examples. The exercises throughout the text are personal involvement opportunities for reflective self-examination and self-discovery. Hamachek views teaching, learning, and growth as personal activities, focusing on human meanings, human understandings, and human experiences.

Don E. Hamachek
Michigan State University

Allyn and Bacon, Inc.
College Division, Department 893
470 Atlantic Ave., Boston, MA 02210

Cultural Democracy, Bicognitive Development, and Education

By MANUEL RAMÍREZ, III
and ALFREDO CASTAÑEDA

All too often educational and mental health institutions in the United States view the cultures of ethnic groups as having a negative effect on a child's intellectual and emotional development. This book shows how American education can better serve all children through the philosophy of "cultural democracy." Specifically focusing on Mexican American children, the authors have studied the variables that affect children's performances in school as well as the development of cognitive styles through socialization. They conclude with many practical suggestions for creating bicultural/bilingual educational environments to stimulate development of cognitive flexibility in children.

October 1974, 189 pp., $11.95/£5.75

Language Use and School Performance

By AARON V. CICOUREL, KENNETH H. JENNINGS, SYBILLYN H. M. JENNINGS, KENNETH C. W. LEITER, ROBERT MacKAY, HUGH MEHAN and DAVID R. ROTH

Here is a significant (and perhaps, disturbing) critique of how elementary school children's abilities, progress, and potential are evaluated. It is based on a thorough investigation of how children and those who evaluate them actually behave and interact. Topics include; Ad Hocing in the Schools: A Study of Placement Practices in the Kindergartens of Two Schools; Accomplishing Classroom Lessons; and Some Basic Theoretical Issues in the Assessment of the Child's Performance in Testing and Classroom Settings.

1975, in preparation
A volume in the Cognitive Sociology Series

The Next Generation

AN ETHNOGRAPHY OF EDUCATION IN AN URBAN NEIGHBORHOOD

By JOHN U. OGBU

This timely volume offers a penetrating analysis of school failure as a historical adaption to unequal opportunity. It probes the forces that created and still sustain this pattern of failure: local social stratification, myths and stereotypes which support that system, the behavior of local school personnel and attitudes of ghetto residents toward competition with members of the dominant society.

1974, 270 pp., $11.50/£5.50
A volume in the Studies in Anthropology Series



Academic Press, Inc.

A Subsidiary of Harcourt Brace Jovanovich, Publishers
111 Fifth Avenue, New York, N.Y. 10003
24-28 Oval Road, London NW1, 7DX

CHARLES C THOMAS • PUBLISHER

EDUCATING EXCEPTIONAL CHILDREN IN A CHANGING SOCIETY *by* **Harold D. Love,** *State College of Arkansas, Conway.* Basically about children who are exceptional, but who still function in our society with the help of many people. Although these children are exceptional, they still have basic human motives, drives, and needs, making their only difference the need for additional help. It has long been the wish of the author that a book be written which could be used for undergraduates and also for beginning graduate students. Most books acceptable for graduate students are too difficult for the undergraduate. This does not have to be, and Doctor Love has proved this by writing a book which is not too difficult for undergraduates, nor too easy for graduate students. The author has tried not only to write an integrated and unified book about exceptional children, but to place the parents of such children and auxiliary personnel and agents in the proper perspective. This makes the emphasis in this text different from most. The tone of complacency in special education is discussed, along with the controversy surrounding methods used in educating exceptional children. This book deals not only with children who have mental and physical problems, but also with children who are intellectually gifted and need special attention to develop their talents. It also deals with the child who has average or above average intelligence, but who still suffers from educational problems. *'74, 264 pp., 19 il., 7 tables, $9.75*

ISSUES AND TRENDS IN BEHAVIOR THERAPY *edited by* **Henry E. Adams,** *Univ. of Georgia, Athens, and* **Irving P. Unikel,** *Veterans Administration Hospital, Atlanta. (10 Contributors)* The purpose of this book is to evaluate current issues and trends in the area of behavior therapy. A paragraph or two of introductory material preceding each chapter will fit its contents into the overall scheme. The beginning chapters are concerned respectively with issues in practice and research and to elaborate upon its aims and purposes. Part I of this book, about 100 pages, is concerned with issues and theoretical developments. Several contributors discuss issues in the practice of behavior therapy in a collaborative effort. Some papers subsequently deal in depth with one major issue and source of criticism for behavior therapy, its capability for dealing with complex cases without specific circumscribed behaviors or symptoms, particularly "existential" problems. Following this, an overview of the research issues in the area are given. These include the general difficulties which accrue when matters are dismissed as being "academic", the need for the study and integration of physiological variables and the necessity for a focus on the development of behavior since the ways maladaptive behaviors are successfully treated are not necessarily indications of the manner in which these behaviors are acquired. The book displays the scope of behavioral treatment techniques by working with obsessive-compulsive disorders. Finally, the evaluation of nonspecific treatment effects is discussed. An original and stimulating work on the influence of demand and contextual variables in behavioral fear assessment. *'73, 288 pp., 38 il., 14 tables, cloth-$10.95, paper-$7.95*

CLASSROOM DISCIPLINE: A Positive Approach *by* **Susan Bray Stainback** *and* **William Clarence Stainback,** *both of Virginia State College, Petersburg.* The primary focus of this book is on straightforward, specific, workable techniques for classroom control. In essence, the authors outline exactly what has to be done in the real, live classroom situation to control disruptive behavior. The book's purpose is to provide prospective as well as practicing classroom teachers, supervisors and principals with techniques that can be used to humanize classroom management. Viable alternatives to outmoded techniques of classroom control, such as threats and corporal punishment, are also presented. The impetus for this book's development was the result of the authors' experiences in working as supervisors of beginning and experienced teachers. The book is divided into four parts. A comparison of the behavior modification strategy adopted by the authors with two other popular strategies (psychodynamic, sensory-neurological) is presented in Part I. Preventive techniques for maintaining classroom discipline as well as measures that can be employed to deal with disruptive behavior after it occurs are outlined in Part II. Part III includes techniques for changing the behavior of extremely disruptive students. In addition, various reinforcers that can be used in the school setting and the ethical issues involved in using reinforcement principles are discussed. This text is an attempt to fill this void in the literature. Selected articles that present ways of preventing and/or dealing with disruptive behavior are provided in Part IV. *'74, 180 pp., $5.50, paper*

AUDITORY PERCEPTUAL DISORDERS AND REMEDIATION *by* **Bernice E. Heasley,** *Human Development and Counseling Associates, North Canton, Ohio. Foreword by* **Kendall K. Ward.** This text is written in clear, concise language with terms being fully explained so the content may be readily understood by both the professional and lay reader. Educators and families of children or adults who exhibit learning or social problems related to defective processing of auditory stimuli will find this book quite useful. Specific educators who will find the book interesting include speech therapists, hearing and language pathologists, audiologists, otologists, special education teachers, and teachers in normal classrooms which contain a percentage of children who exhibit auditory perceptual disorders in varying degrees. Based on the Developmental Model of Speech and Language, the book provides auditory perceptual training lessons which follow a carefully structured hierarchy in which lessons are graduated in terms of difficulty, length of lesson time, and numbers of processes to be introduced. This structure closely approximates the normal sequential landmarks of childhood development in the areas of hearing, language, and speech. Remediation tasks are based on common environmental and speech sounds and require no expensive equipment. Topics include components of auditory processing and criteria for recognizing symptoms of auditory perceptual problems. *'74, 128 pp., 20 il., 1 table, $8.95*

Orders with remittance sent, on approval, postpaid

301-327 EAST LAWRENCE AVENUE • SPRINGFIELD • ILLINOIS • 62717